Predicció de vendes amb sèries temporals

Autora: Anna Batlle Olmo, 1363013
Tutor: Josep Lluís Solé
Grau d' Estadística Aplicada (UAB)

Índex

Introducció

Vaig realitzar les pràctiques externes curriculars a Whirlpool S.A., empresa multinacional d'electrodomèstics. La meva funció allà era la realització de la predicció de vendes mensuals dels seus productes mitjançant la utilització dels seus mètodes propis. L'objectiu d'aquest treball és fer la predicció de vendes d'aquests mateixos productes, però amb els mètodes apressos durant els estudis de grau.

En primer lloc, es realitzarà un anàlisi global de les dades, agafant els productes en conjunt i sense dividir per cap variable, per veure el seu desenvolupament al llarg del temps, i s'estudiarà el mètode de predicció utilitzat per l'empresa. A continuació, es procedirà a l'anàlisi específic per categoria de producte. En aquest apartat, s'estudiarà la possible dependència temporal de les dades. Segons els resultats obtinguts, s'intentarà trobar models que ajustin les dades i serveixin per fer les prediccions dels següents dotze mesos.

Dades

La base de dades conté valors **mensuals des de gener de 2015 fins a gener de 2017**. Cal destacar que només dos anys de dades mensuals és una base de dades molt petita, però donat que és el que tenen, és el que he utilitzat. Les variables utilitzades són:

Demanda. Variable numèrica discreta que conté la suma de les vendes normals i les vendes produïdes en període promocional.

Promoció. Variable binària que indica si durant un més concret hi va haver promoció, amb el valor 1, o si no hi va haver promoció, amb el valor 0.

Categoria. Variable categòrica nominal que explica la subdivisió de la tipologia dels productes. La base de dades conté un total de 32 categories que són les següents: Acc (accessoris), Air (aire condicionat), BIFP (preparació de menjar d'encast), BIFreez (congeladors d'encast), BIRefr (neveres d'encast), BIRefrFreez (combinació de nevera i congelador d'encast), Coffee (cafeteres), Combi (combinació de nevera i congelador), Cooker (cuines), Dish (rentavaixelles), Dryers (secadores), FCPAcc (accessoris de preparació de menjar), FP (preparació de menjar), FPPAcc (accessoris de preparació de menjar, altres), Freez (congeladors), FSFreez (congeladors aïllats), FSIceMak (geladores aïllades), FSRefr (neveres aïllades), FSRefrFreez (combinació de nevera i congelador aïllats), FSSPAcc (accessoris de preparació de menjar aïllats), HCPAcc (accessoris, altres), Hob (plaques de cuina), Hood (campanes extractores), Juice (exprimidors), MWO (microones), OutFP (productes d'exterior), Oven (forns), Refr (neveres), Toaster (torradores), WashDry (combinació de rentadora i secadora), Washer (rentadores), WasteM (eliminació de brossa).

Mes. Variable categòrica ordinal que pren valors des de 1501 (gener de 2015) fins a 1701 (gener de 2017), on les dues primeres xifres indiquen l'any i les altres dues indiquen el mes.

Trimestre. Variable fictícia categòrica ordinal que indica l'estació de l'any. Està creada a partir de la variable *mes* i es composa de: 1 per l'hivern (desembre, gener i febrer), 2 per la primavera (març, abril i maig), 3 per l'estiu (juny, juliol i agost) i 4 per la tardor (setembre, octubre i novembre).

Anàlisi global

Com a primer anàlisi, es realitza un gràfic de comparació entre les vendes normals i les produïdes per promocions, bàsicament per veure si aquestes tenen un volum a considerar.

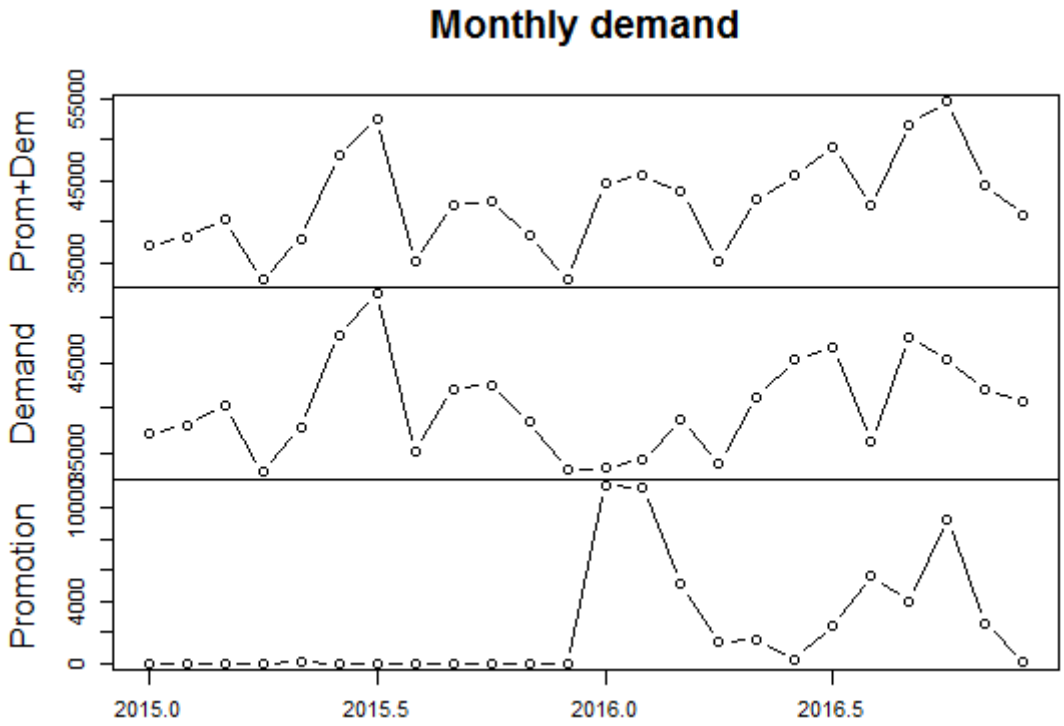

Figura 1. De baix a dalt: vendes en promoció, vendes sense promoció, i suma de vendes amb i sense promoció, al llarg dels 2 anys d'estudi.

Tal i com es pot observar a la *Figura 1*, sembla ser que les promocions no es comptabilitzaven gaire abans de l'any 2016, així que es podria entendre que algunes deuen estar incloses en les vendes genèriques (*demand*). Si es mira la demanda total, sembla que podria haver una recursivitat que consisteix en un pic que cau, després torna a pujar i torna a caure, etc.

A la *Figura 2* es pot observar clarament la similitud de la demanda al llarg de l'any, entre els dos anys d'estudi. Tot i que les proporcions siguin diferents i l'empresa no tingui una explicació raonada d'aquest comportament, resulta evident la semblança de les seves formes.

No obstant, tenint en compte que només hi ha dades de 2 anys, si es fes un model que tingués en compte aquesta recursivitat, només tindria 2 dades per cada mes, el qual produiria una estimació de prediccions no gaire millor que un simple model aleatori.

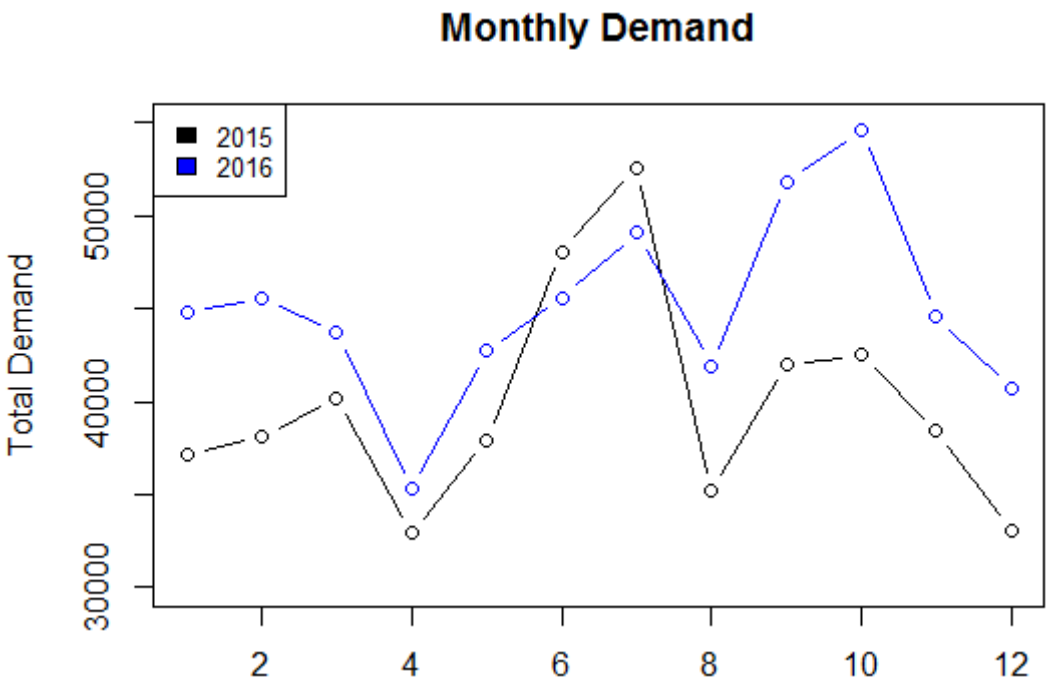

Figura 2. Gràfics de la demanda total anual.

Tendència i recursivitat

Una teoria bàsica d'ajust de models a sèries temporals és veure si la tendència i la recursivitat són suficients per explicar les dades. La tendència ve definida pel pendent de la recta de regressió que ajusta les dades pel temps. La recursivitat és el punt mig per cada moment repetit, és a dir, en aquest cas la mitjana dels geners, la dels febrers, etc. Si fossin suficients per explicar la demanda, els residus del model haurien de tindre forma de soroll blanc.

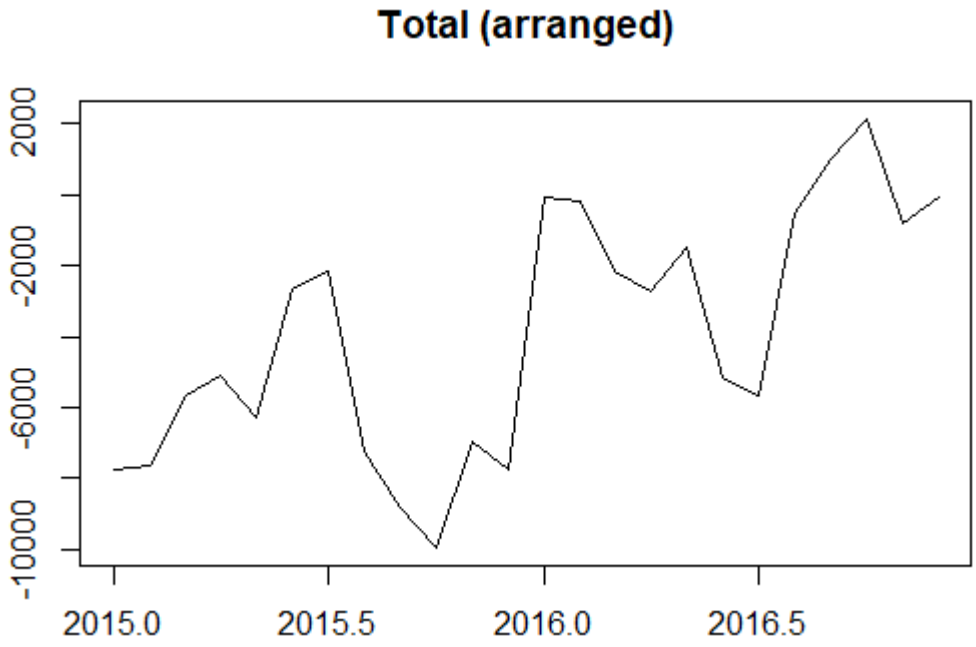

Figura 3. Gràfic de la demanda total sense tendència ni recursivitat.

La *Figura 3* mostra les dades resultants de treure la recursivitat i la tendència a les dades. Tot i així, no semblen tenir aspecte de dades aleatòries, que seria el que hauria de passar.

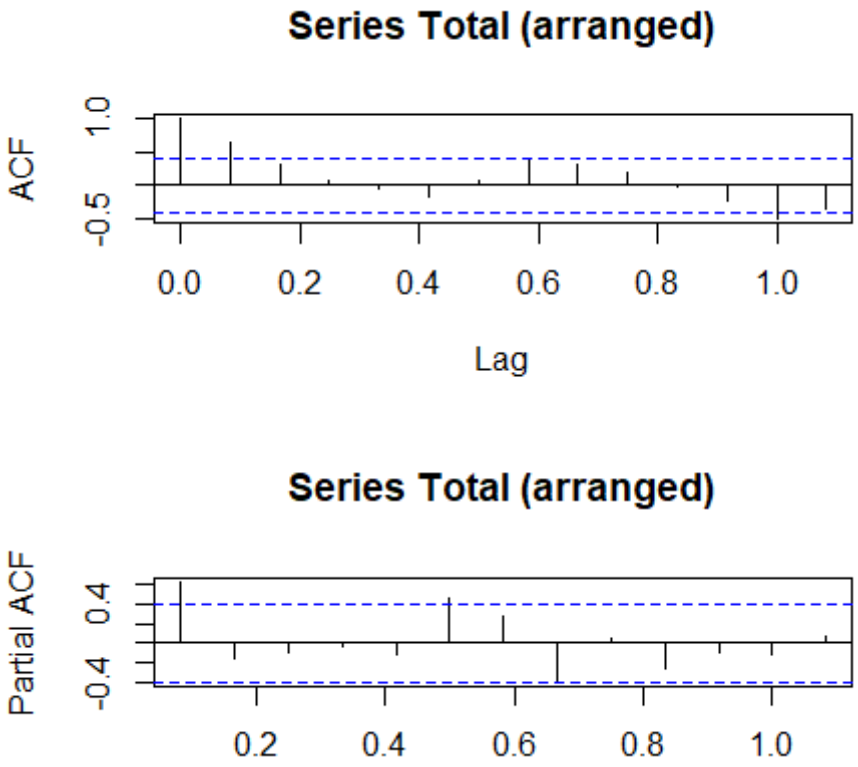

Figura 4. Gràfics d'auto correlació i d'auto correlació parcial de la demanda sense tendència ni recursivitat.

En la *Figura 4* es pot observar les seves funcions d'auto correlació i auto correlació parcial. Tampoc semblen ser soroll blanc. En tot cas, per més seguretat, es pot fer un test d'independència, per exemple, el de Box-Pierce[1]. El resultat confirma que hi ha auto correlació, per tant, el model bàsic proposat no és suficient per fer un bon ajust de les dades. Tenint això en compte, a continuació cal analitzar més detingudament la independència de la demanda.

Independència

El primer pas per analitzar una sèrie temporal consisteix en comprovar que les dades estiguin auto correlacionades i, per tant, es pugui produir un model que expliqui el futur a partir del passat, ja que en cas de no ser-ho, es considera que és soroll blanc, és a dir, el futur no depèn del passat sinó que és aleatori.

Hi ha dues maneres de realitzar aquesta comprovació: una és observar les funcions d'auto correlació i d'auto correlació parcial, i l'altra és fer un test d'independència (per exemple el test de Box-Pierce, que té com a hipòtesi nul·la la independència de la sèrie temporal).

[1] Box-Pierce test: X-squared = 26.7296, df = 12, p-value = 0.00845

En les funcions d'auto correlació es mostra la importància de cada retard sobre la dada actual. Les franges blaves són les bandes de significació, és a dir, si una línia va més enllà de las línies blaves és perquè la dada d'aquell retard té influència en la dada actual, amb un 95% de confiança (hi ha un 5% de les vegades que una línia sobresurt però realment no és important).

El test d'independència de Box-Pierce fa el mateix, però des d'un punt de vista numèric, i no gràfic, i de manera més general. No compara un per un tots els retards, sinó que busca si hi ha algun que sigui important a l'hora d'explicar la dada actual, sent la hipòtesi nul·la que no hi ha cap. Si el p-valor fos inferior a 0.05 , es rebutjaria aquesta hipòtesi i es diria que sí que hi ha algun retard que influeix en la dada actual, amb un 95% de confiança.

El més adequat és, per més seguretat, fer les dues coses per tal de reafirmar la seva conclusió.

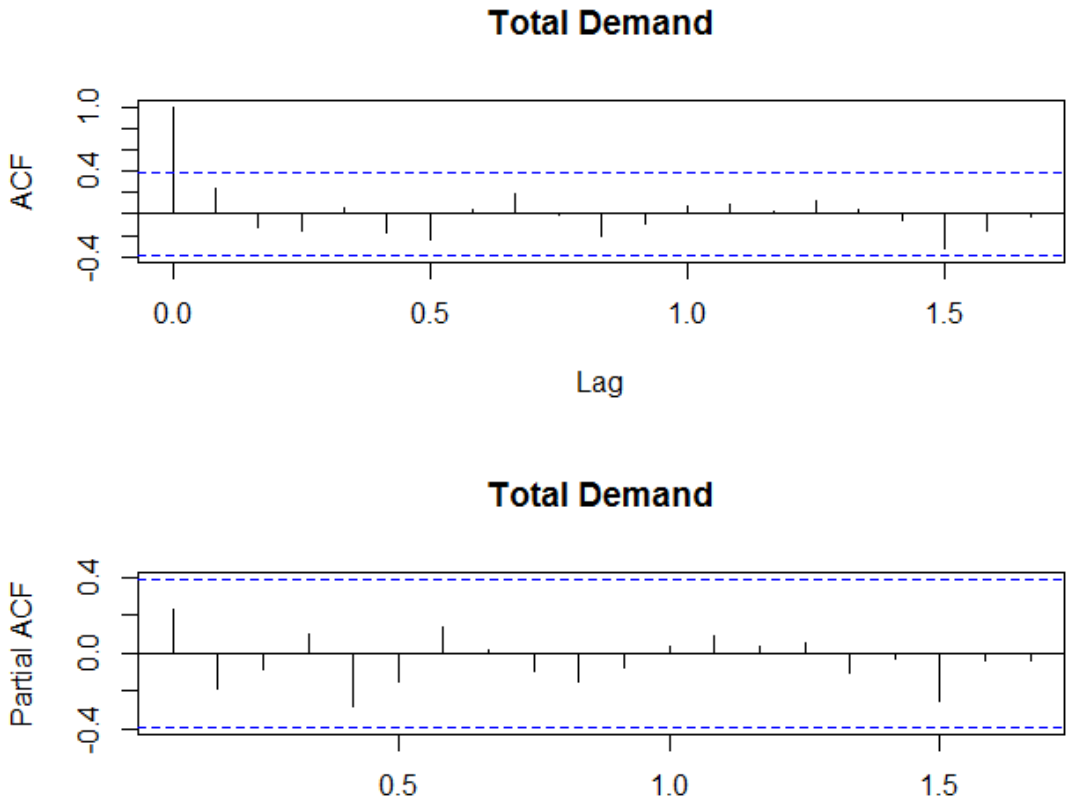

Figura 5. Gràfics d'auto correlació i d'auto correlació parcial de la demanda.

Com es pot observar, la *Figura 5* mostra que no hi ha auto correlació en cap retard i el test d'independència[2] no troba mostres estadísticament significatives de que les dades siguin dependents entre elles. Així doncs, la demanda global sembla ser soroll blanc.

[2] Box-Pierce test: X-squared = 1.8024, df = 1, p-value = 0.1794

Model a 3 mesos

El model que utilitzaven a l'empresa (*Model 1*) era aproximadament un model de sèries temporals amb tres retards, ja que no deixa de ser el passat com a variables explicatives, cadascuna amb el seu respectiu coeficient (*Model 2*).

$$Dem_t = pes_0 \left(0.5 \cdot \frac{1}{pes_1} \cdot Dem_{t-1} + 0.3 \cdot \frac{1}{pes_2} \cdot Dem_{t-2} + 0.2 \cdot \frac{1}{pes_3} \cdot Dem_{t-3}\right) \quad (1)$$

$$Dem_t = \alpha_1 Dem_{t-1} + \alpha_2 Dem_{t-2} + \alpha_3 Dem_{t-3} \quad (2)$$

on $\alpha_1 = \frac{pes_0}{pes_1} \cdot 0.5$, $\alpha_2 = \frac{pes_0}{pes_2} \cdot 0.3$ i $\alpha_3 = \frac{pes_0}{pes_3} \cdot 0.2$.

Els pesos als que fan referència els models són el resultat de mirar la desviació de la demanda del mes en qüestió respecte la mitjana de demanda dels últims 12 mesos. És a dir, per exemple, si es vol predir gener de 2017, seria agafar la demanda del gener de 2016 i dividir-la per la demanda mitjana dels últims 12 mesos (del desembre de 2015 fins el desembre de 2016). La fórmula seria la següent:

$$pes_0 = \frac{Dem_{t-12}}{\overline{Dem}_{\{t-12:t-1\}}}$$

En el nostre cas, simplement es realitza una estimació automàtica dels coeficients amb un model lineal (*Model 3.1* i *3.2*).

$$Dem_t = \beta_0 + \beta_1 Dem_{t-1} + \beta_2 Dem_{t-2} + \beta_3 Dem_{t-3} + u_t \quad (3.1)$$

Com que les dades són soroll blanc, el model creat amb els retards no hauria de ser gaire més útil que fer una estimació aleatòria. Per comprovar-ho, cal fer el model lineal equivalent al que fan servir, la sortida d'R del qual és la següent:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	35131.5457	4999.8131	7.027	8.14e-07	***
Dem1	0.1945	0.1555	1.251	0.226	
Dem2	-0.1277	0.1614	-0.791	0.438	
Dem3	0.1187	0.1265	0.939	0.359	

Residual standard error: 5951 on 20 degrees of freedom
Multiple R-squared: 0.1375, Adjusted R-squared: 0.008097
F-statistic: 1.063 on 3 and 20 DF, p-value: 0.3871

I equival al model:

$$Dem_t = 35131.5457 + 0.1945 Dem_{t-1} - 0.1277 Dem_{t-2} + 0.1187 Dem_{t-3} \quad (3.2)$$

Cap de les variables és estadísticament significativa, a més, el model en sí tampoc ho és ja que el p-valor del *F-statistic* és molt més gran de 0.05. Per si fos poc, té un R^2 ajustat molt proper a zero, el qual significa que és igual de útil la informació que pugui proporcionar el model que fer un procediment aleatori. En definitiva, el model utilitzat d'aquesta manera no serveix per res, tot i que a l'empresa li pugui resultar útil en el seu procés propi de predicció de vendes.

Anàlisi per categories

Precisament perquè analitzar les dades en el seu conjunt no té gaire sentit al no poder modelar-se, és el motiu pel qual es necessari prendre un nivell inferior per l'anàlisi, que seria la divisió per categoria de producte. Per tindre una visió general del funcionament de cada categoria, la *Figura 6* mostra un gràfic de caixa de la demanda per cada una d'elles.

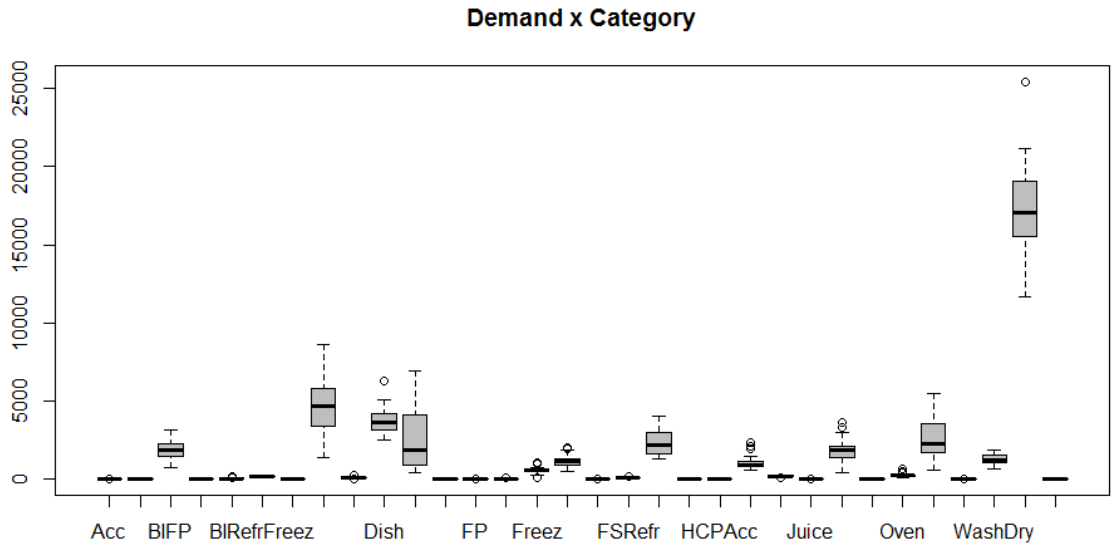

Figura 6. Gràfics de caixa de la demanda total per categoria.

El que es pot observar d'aquest gràfic és que hi ha una categoria (*Washer*) que destaca molt per sobre de la resta (de fet, les rentadores són el producte estrella de l'empresa) i que hi ha moltes que o bé no venen gaire o bé no es pren nota de les vendes. És possible que, en part, l'anàlisi global no tingués sentit a causa de totes aquestes categories, que involuntàriament creen un biaix. A més, cal destacar la gran diferència entre variabilitats i que algunes categories es mantenen molt més estables que altres.

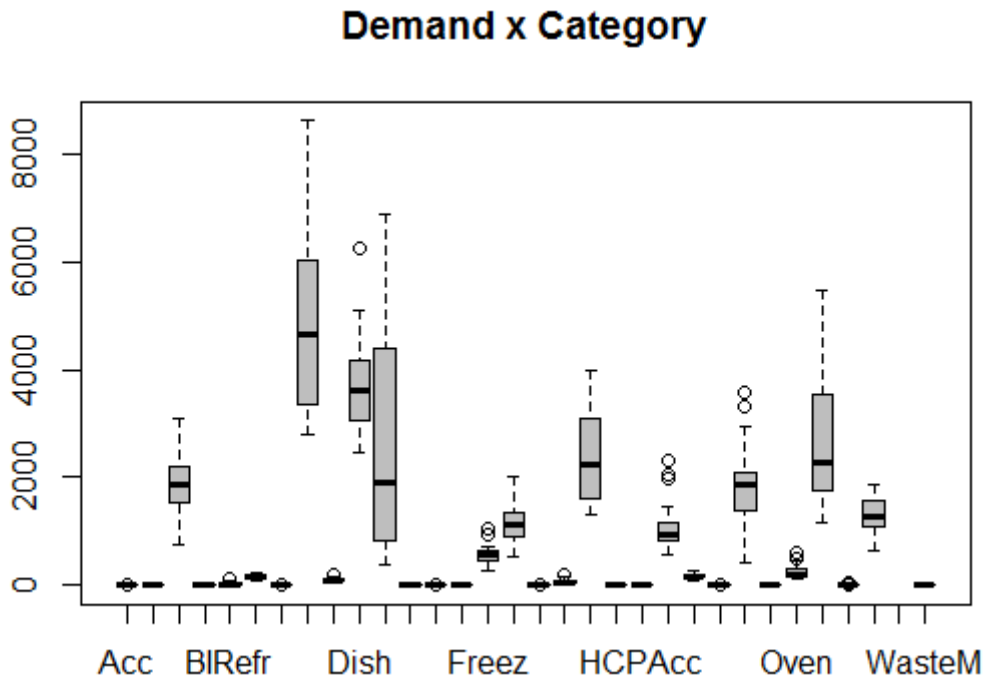

Figura 7. Gràfics de caixa de la demanda total per categoria,
en categories amb demanda inferior a 10 000 unitats.

La *Figura 7* és el mateix gràfic de caixa que la *Figura 4* però excloent la categoria *Washer* per tal de poder observar més clarament la variabilitat de la resta de les categories. Sembla ser que la variabilitat també és elevada en moltes altres categories com per exemple, *Combi* o *Dryers*.

Independència

De la mateixa manera que abans, cal veure si les dades són independents entre elles o si mostren auto correlació. En aquest cas, però, degut a l'elevat nombre de categories, es realitzen tots els gràfics d'auto correlació i tests d'independència, però sense mostrar-los.

La majoria de les categories mostren gràfics amb el mateix aspecte que la *Figura 8*, amb la confirmació dels tests corresponents de que efectivament no mostren signes de correlació[3]. Per tant, la demanda futura de la majoria de les categories no depèn de la demanda passada.

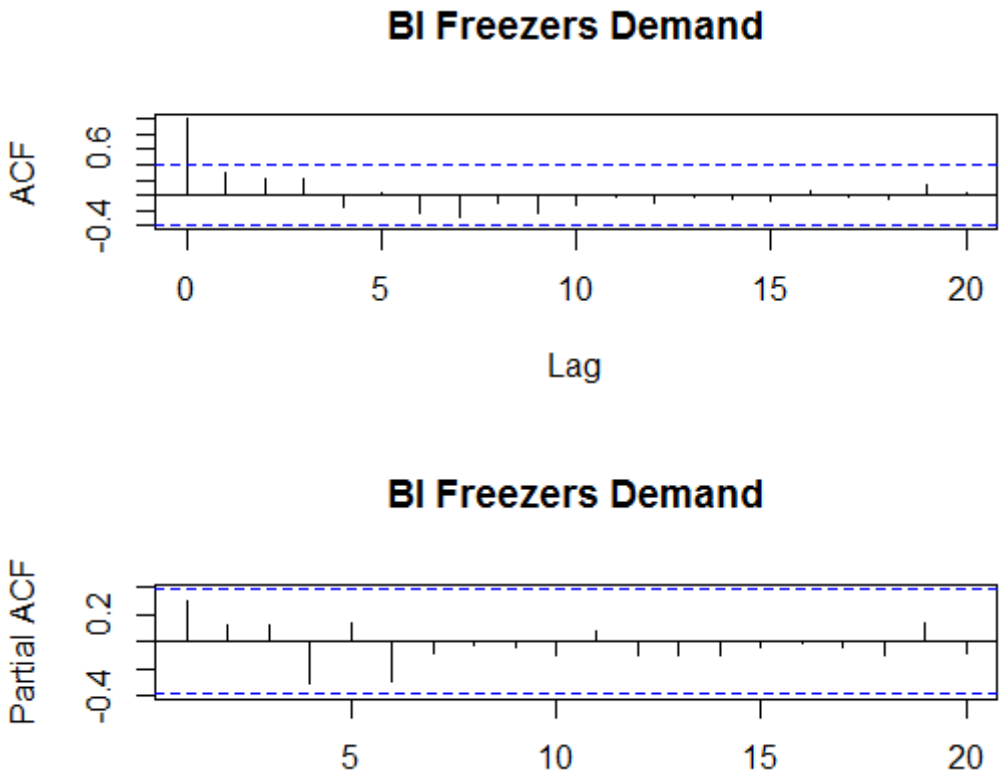

Figura 8. Gràfics d'auto correlació i d'auto correlació parcial dels congeladors d'encast, com a representació de les categories sense auto correlació.

Les categories que han mostrat evidències significatives d'auto correlació al 95% de confiança al test, amb els seus respectius p-valors, són: *Combi* – 0.001474, *Cooker* – 0.044807, *Dryers* – 0.000516, *Freez* – 0.014678 i *Refr* – 0.000774 . Aquestes categories s'estudiaran més endavant amb deteniment per trobar d'ajustar-ne models autoregressius que serveixin posteriorment per fer les prediccions de la seva demanda.

[3] Box-Pierce test: X-squared = 2.4599, df = 1, p-value = 0.1168 (resultat del test dels congeladors d'encast, com a representació de resultat obtingut a les categories sense auto correlació)

Correlació entre les categories

Aquestes cinc categories tenen dependència del seu passat, però potser també tenen dependència entre elles. La seva matriu de correlacions és la següent:

	Combi	Cooker	Dryers	Freez	Refr
Combi	1.0000000	0.3728090	-0.6950307	0.5840855	0.9001917
Cooker	0.3728090	1.0000000	-0.2959616	0.6634017	0.3694639
Dryers	-0.6950307	-0.2959616	1.0000000	-0.4065856	-0.7693268
Freez	0.5840855	0.6634017	-0.4065856	1.0000000	0.5655234
Refr	0.9001917	0.3694639	-0.7693268	0.5655234	1.0000000

Sembla ser que hi ha una elevada correlació entre les categories *Combi* i *Refr*, i una correlació negativa considerable en les combinacions de *Dryers* amb *Combi* i *Refr*. La relació entre *Combi* i *Refr* és lògica, ja que els dos són tipus de neveres i, per tant, és normal que les dues tinguin el mateix perfil de venda. La relació inversa entre aquestes categories i les assecadores, podria ser deguda, per exemple, a que aquestes es venen menys a l'estiu, que és quan fa més calor que és quan es venen més neveres degut a que amb l'augment de calor cal augmentar la potència de refredament i és més fàcil que es produeixin averies i, per tant, hi hagi un increment en la demanda de neveres. No és una relació causal (en principi les neveres i les assecadores no tenen res a veure entre elles) però sí una relació casual a través de la variable temporal.

Per aprofitar la informació proporcionada per aquestes correlacions elevades, es pot realitzar un anàlisi de components principals. L'objectiu d'aquest anàlisi és fer transformacions lineals de les dades tal que la variància més gran del conjunt de dades queda recollida en la primera component principal, la segona variància més gran en la segona component principal, etc. Amb això s'aconsegueix crear noves "variables" no correlacionades entre elles que expliquin més clarament la seva influència sobre la variable resposta.

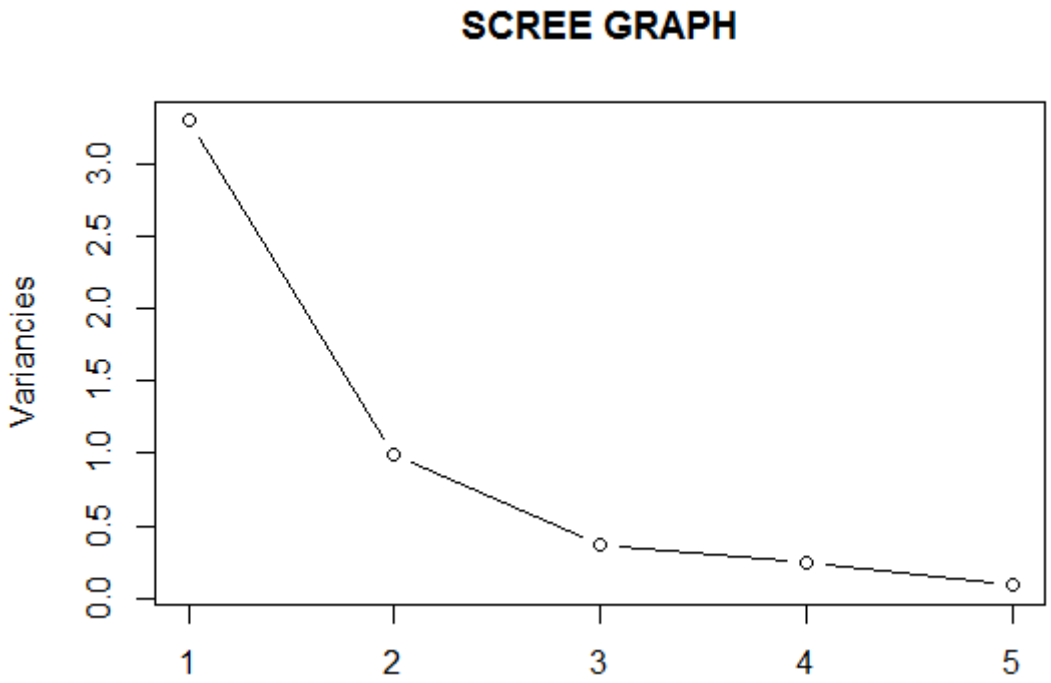

Figura 9. Scree Graph de components principals.

Tant pel mètode de la proporció de les variàncies (escollir suficients per explicar com a mínim el 80% de la variància: *Cumulative proportion > 0.8*) [4] com pel Scree Graph (*Figura 9*) (escollir tantes com punts hi hagi abans d'arribar a la zona plana del gràfic), el resultat és que calen dues components principals. Els vectors propis de les components principals per cada categoria són els següents:

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Combi	-0.498	0.246	0.349	0.414	0.632
Cooker	-0.344	-0.698	-0.482	0.403	
Dryers	0.441	-0.389	0.639	0.469	-0.162
Freez	-0.429	-0.467	0.464	-0.618	
Refr	-0.505	0.290	0.148	0.253	-0.758

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
SS loadings	1.0	1.0	1.0	1.0	1.0
Proportion Var	0.2	0.2	0.2	0.2	0.2
Cumulative Var	0.2	0.4	0.6	0.8	1.0

Les components principals són com variables inherents que no es poden observar directament. L'única manera d'arribar a elles és mitjançant la combinació de variables que sí es poden observar, en aquest cas, les categories. El significat dels vectors propis no és més que una forma d'expressar quina proporció de cada categoria hi ha a cada component. Per exemple, en aquest cas, prenent la primera component, es podria dir que hi ha una variable no observable directament que és:

$$cp_1 = -0.498\, Combi - 0.344\, Cooker + 0.441\, Dryers - 0.429\, Freez - 0.505\, Refr$$

En aquesta, apareixen totes les categories i en proporcions semblants, així que no té una interpretació clara a simple vista, tot i que sí que en deu tindre. En canvi, mirant l'última component principal, que és:

$$cp_5 = 0.632\, Combi - 0.162\, Dryers - 0.758\, Refr$$

es pot observar que les que més pes tenen són les categories *Combi* i *Refrigerators*, mentre que *Dryers* influeix molt poc i les altres ni tan sols apareixen. Això podria voler dir que hi ha algun factor en principi desconegut que relaciona fortament aquestes dues categories i que, alhora, les diferencia de la resta, com podria ser la teoria anterior sobre el pla de vendes ajustat a l'estació de l'any.

Tot i que no tinguin gaire funció explicativa donada la seva naturalesa, les components principals resulten útils a l'hora de fer prediccions perquè afegeixen aquesta informació rellevant sobre agrupacions que no es podria obtindre de cap altra manera.

Els següents gràfics mostren les quatre categories (*Figura 10*) i els mesos, de l'1 (gener de 2015) al 24 (desembre de 2016), (*Figura 11*) respecte de les dues primeres components principals.

[4] Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	1.8141491	0.9970509	0.60863115	0.50325167	0.30175859
Proportion of Variance	0.6582274	0.1988221	0.07408638	0.05065245	0.01821165
Cumulative Proportion	0.6582274	0.8570495	0.93113590	0.98178835	1.00000000

Mirant només el comportament de la primera component principal al llarg del temps s'observa una tendència significativament[5] decreixent (*Figura 12*), que es pot utilitzar per fer prediccions del futur (*Figura 13*).

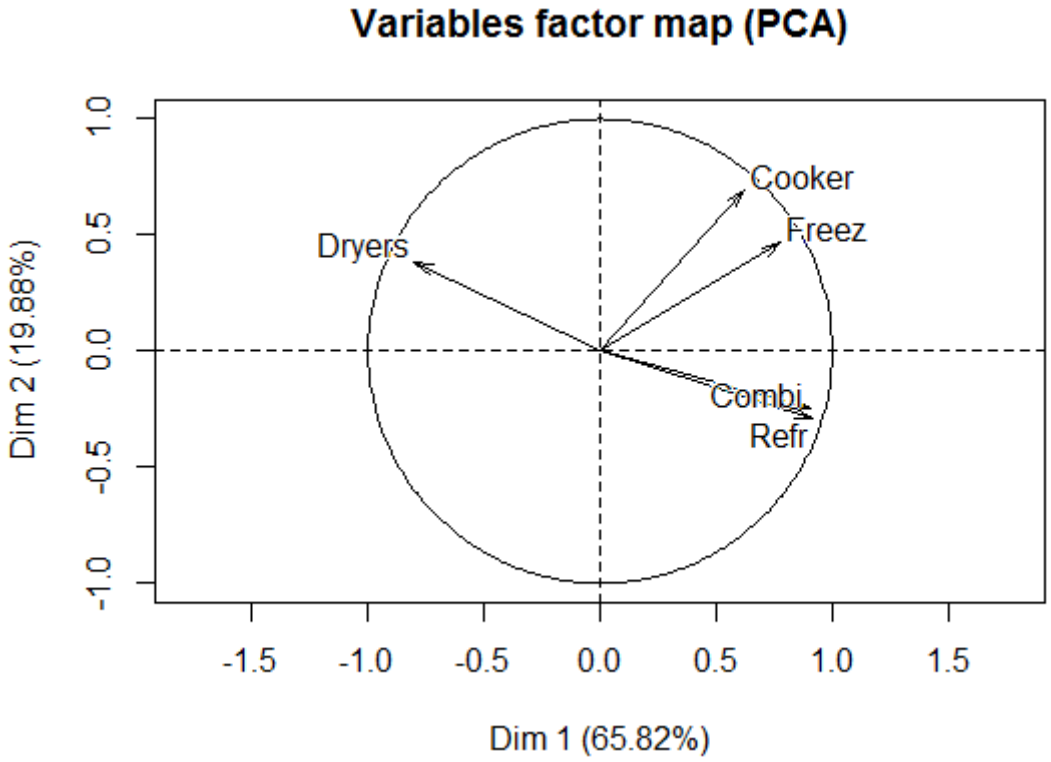

Figura 10. Gràfic de les components principals per categoria.

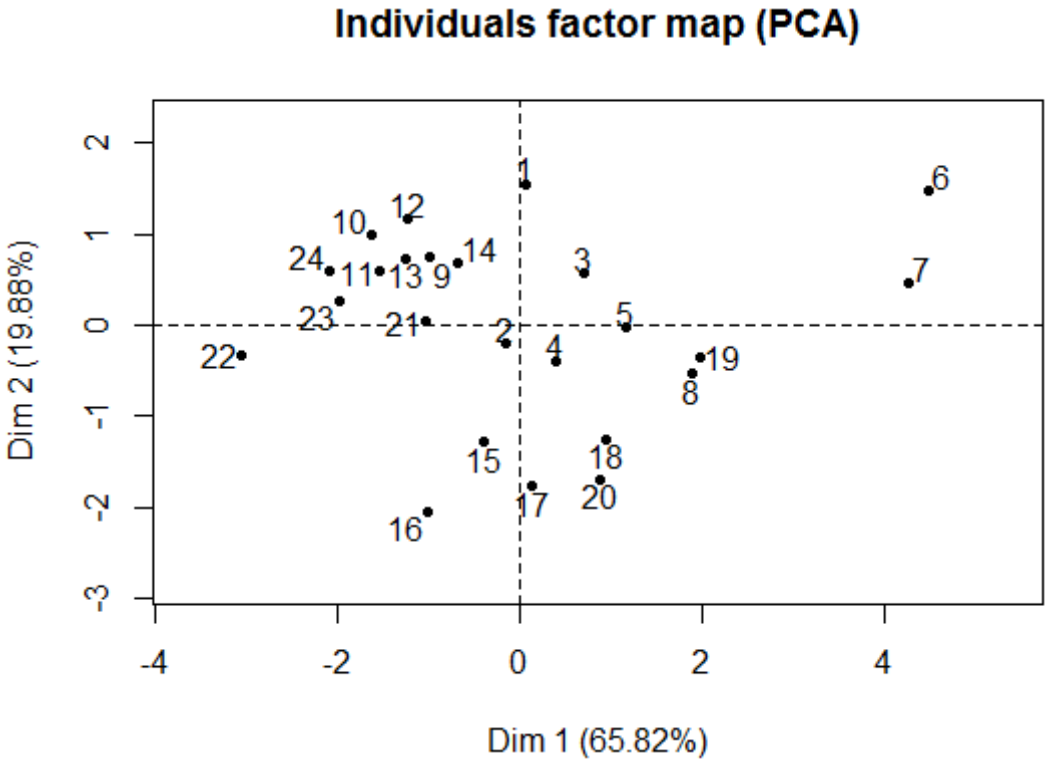

Figura 11. Gràfic de les components principals per mes.

[5]

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2774.8260	1215.3664	2.283	0.0324	*
temp	-1.3764	0.6029	-2.283	0.0324	*

Residual standard error: 1.704 on 22 degrees of freedom
Multiple R-squared: 0.1916, Adjusted R-squared: 0.1548
F-statistic: 5.213 on 1 and 22 DF, p-value: 0.03244

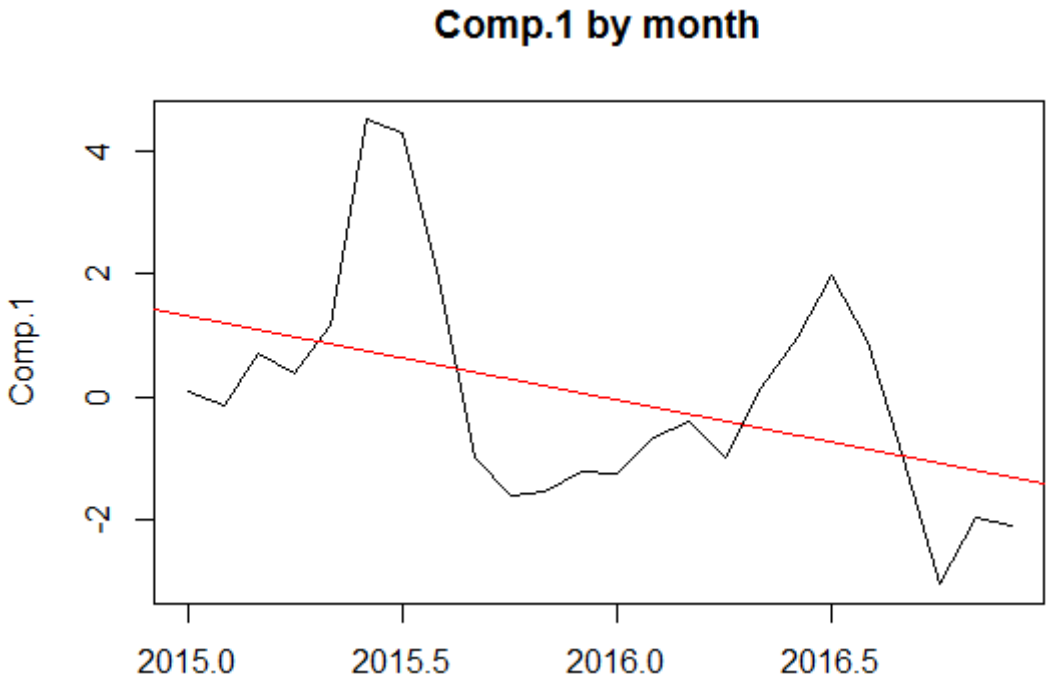

Figura 12. Gràfic de la primera component principals per mes.

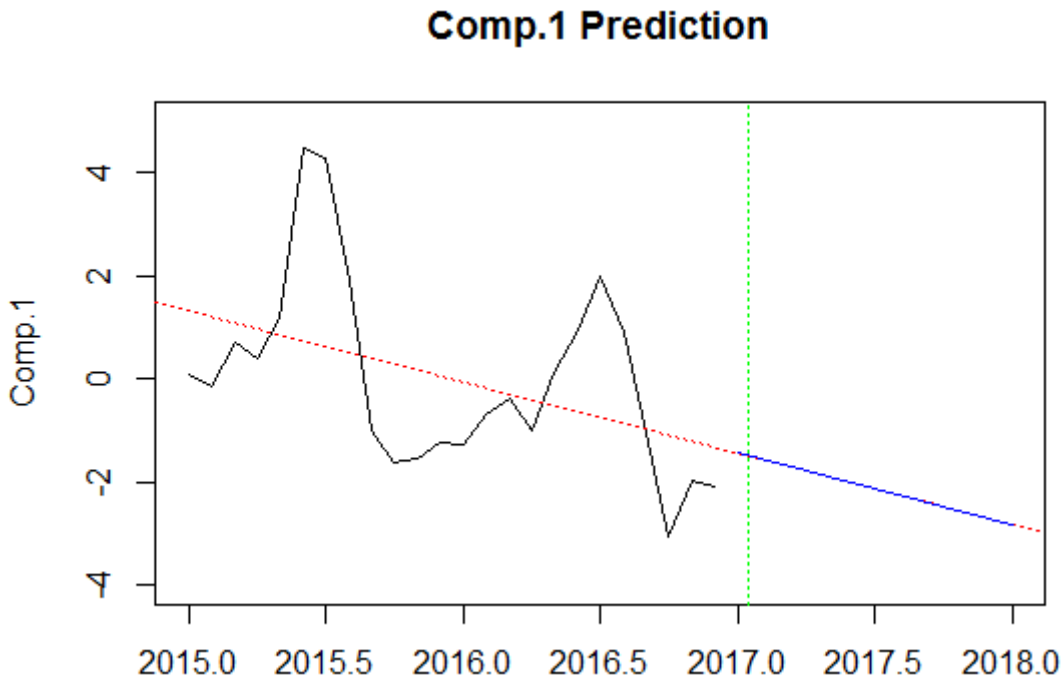

Figura 13. Predicció de la primera component principal pel proper any.

El mateix passa amb la segona component principal, tot i que és una mica menys significativa[6] (*Figura 14* i *15*).

[6]

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1383.1690	681.8551	2.029	0.0548 .
temp	-0.6861	0.3382	-2.029	0.0548 .

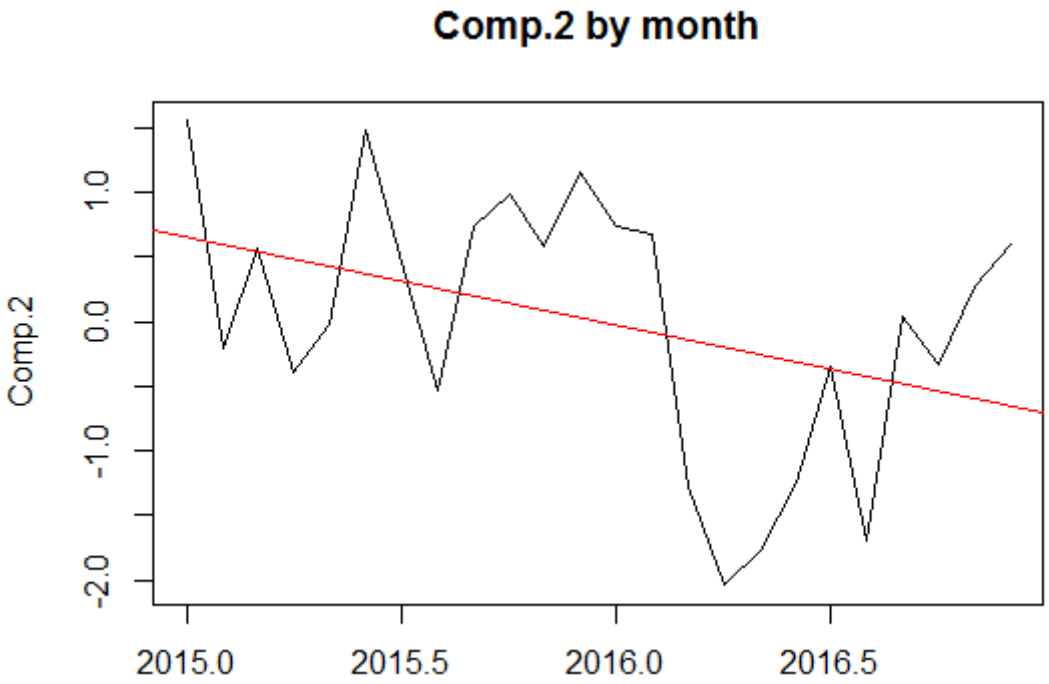

Figura 14. Gràfic de la segona component principal per mes.

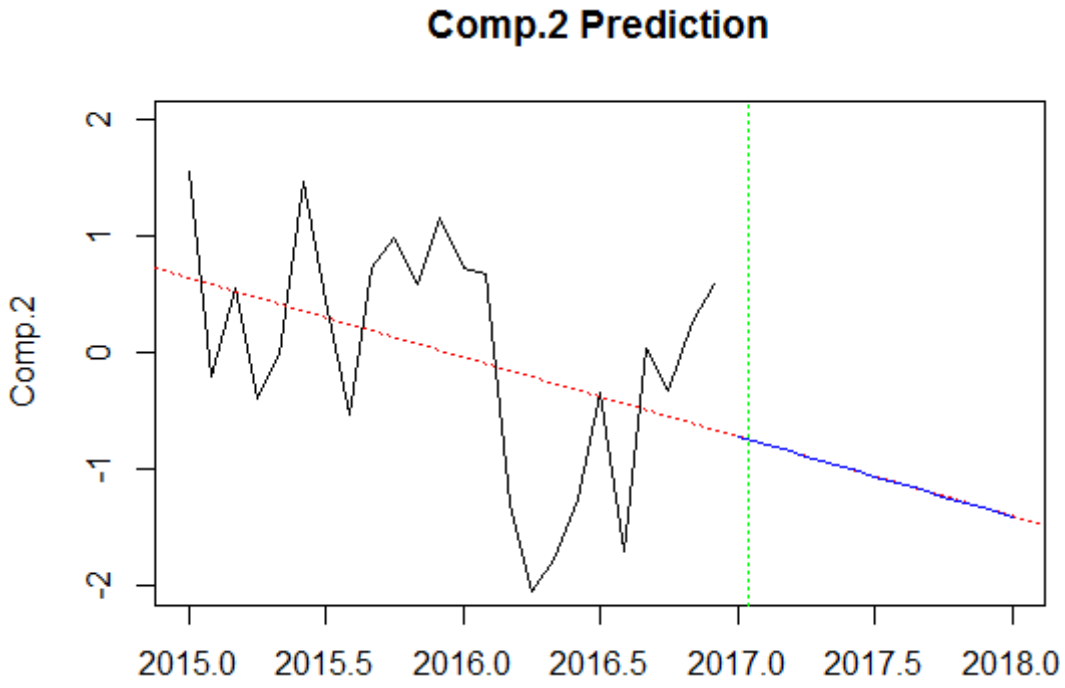

Figura 15. Predicció de la segona component principal pel proper any.

Residual standard error: 0.9558 on 22 degrees of freedom
Multiple R-squared: 0.1576, Adjusted R-squared: 0.1193
F-statistic: 4.115 on 1 and 22 DF, p-value: 0.05478

Models autoregressius

Prenent aquestes quatre categories, es poden trobar models autoregressius per cada una d'elles que ajustin les sèries de dades.

Una sèrie estacionaria és aquella on ni l'esperança ni la variància depenen del temps i la covariància es manté constant pels mateixos períodes de temps, alhora que té un residu aleatori (Soroll Blanc). Els models *ARMA(p,q)* (*AutoRegressive Moving Average*) són els que defineixen un valor present de la sèrie a partir dels *p* moments anteriors i on l'error no és només el del present, sinó que hi influeixen els errors dels *q* moments anteriors.

$$X_t = \alpha + \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}$$

S'ha de complir que $|\phi| < 1$. El valor de α és l'esperança en el cas dels models MA, i una part d'aquesta en els models amb AR. Els valors de p i q es reconeixen a partir dels gràfics d'autocorrelació i d'autocorrelació parcial de la sèrie, veient quins retards són estadísticament significatius en quant a importància a l'hora d'explicar el valor present. Una vegada ajustat un model ARMA, els residus no poden mostrar autocorrelació en cap retard, ja que han de ser aleatoris i no poden tindre relació entre ells.

Combi (combinació de nevera i congelador)

La *Figura 16* mostra la demanda d'aquesta categoria al llarg dels 2 anys.

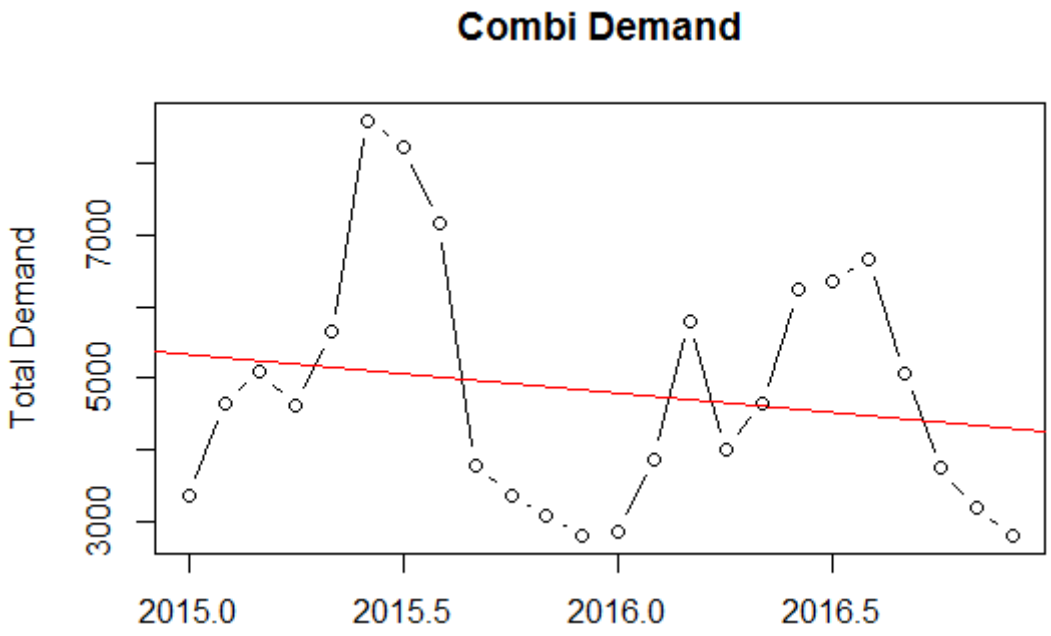

Figura 16. Gràfic de la demanda de Combi al llarg del temps.

Com era d'esperar, sembla ser que les vendes augmenten als mesos d'estiu i disminueixen a l'hivern, el qual genera una estacionalitat en la sèrie. En general, sembla que aquesta categoria ha tingut un descens en les vendes al llarg d'aquest temps, però el coeficient[7] no és significatiu, així que en realitat pot considerar-se constant.

[7]

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1107845.2	1221166.6	0.907	0.374
t	-547.1	605.7	-0.903	0.376

Residual standard error: 1712 on 22 degrees of freedom
Multiple R-squared: 0.03576, Adjusted R-squared: -0.00807
F-statistic: 0.8159 on 1 and 22 DF, p-value: 0.3762

Podria pensar-se que també hi ha certa recurrència anual, és a dir, al comportament al llarg de l'any es repeteix cada any (tal com mostra la *Figura 17*).

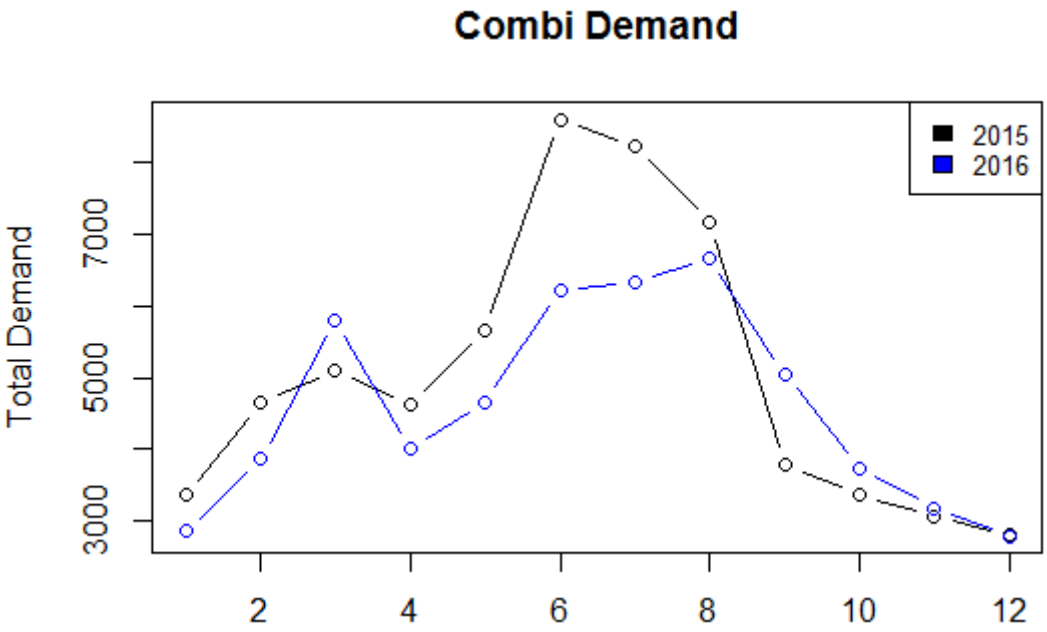

Figura 17. Gràfics de la demanda anual de Combi.

La *Figura 18* mostra la gràfica de les funcions d'auto correlació i d'auto correlació parcial. Pel que s'observa, sembla ser que un model AR(2) seria suficient per ajustar les dades. A més, com a comprovació, es pot utilitzar la funció *auto.arima* que calcula l'AIC de tots els possibles models i els compara per donar automàticament el que mostra un millor ajust.

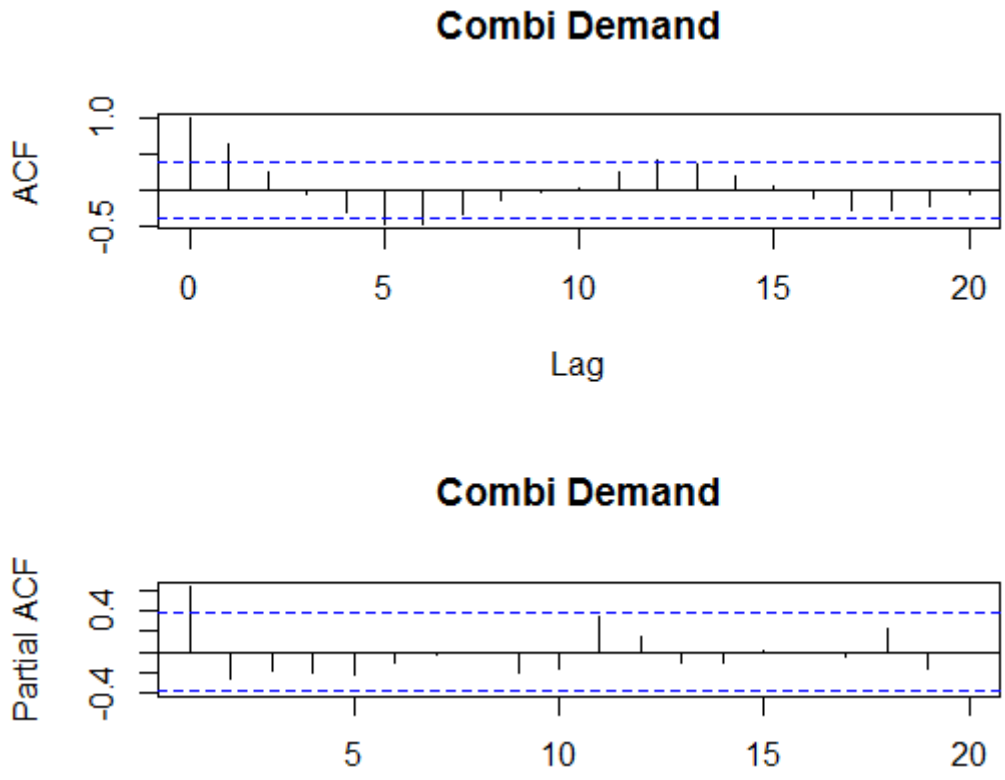

Figura 18. Gràfics d'auto correlació i d'auto correlació parcial de Combi.

El model estimat[8], per tant, seria el següent:

$$X_t = 2074.9948 + 0.9967X_{t-1} - 0.4299\,X_{t-2} + Z_t$$

A la *Figura 19* es pot veure un anàlisi dels residus del model escollit. En el primer gràfic (residus normalitzats) tots s'han de trobar dins el rang entre -2 i 2 perquè pugui considerar-se que el model ajusta bé per totes les observacions, sense que cap destaqui ni quedi fora de lloc. En el segon gràfic (correlograma dels residus) ha de mostrar un valor elevat en 0 i valors petits (dins les bandes de confiança) per la resta de retards, ja que això significa que els residus són Soroll Blanc i, per tant, són realment aleatoris. En el tercer gràfic (p-valors dels retards) es veu el p-valor d'una prova de significança dels retards dels residus, és a dir, si mostren auto correlació en algun punt; que estiguin tots fora de rang és signe d'un bon ajust del model escollit.

En aquest cas, es compleixen tots els requisits, per tant, es pot considerar que el model ajusta bé les dades i és apropiat per la seva utilització.

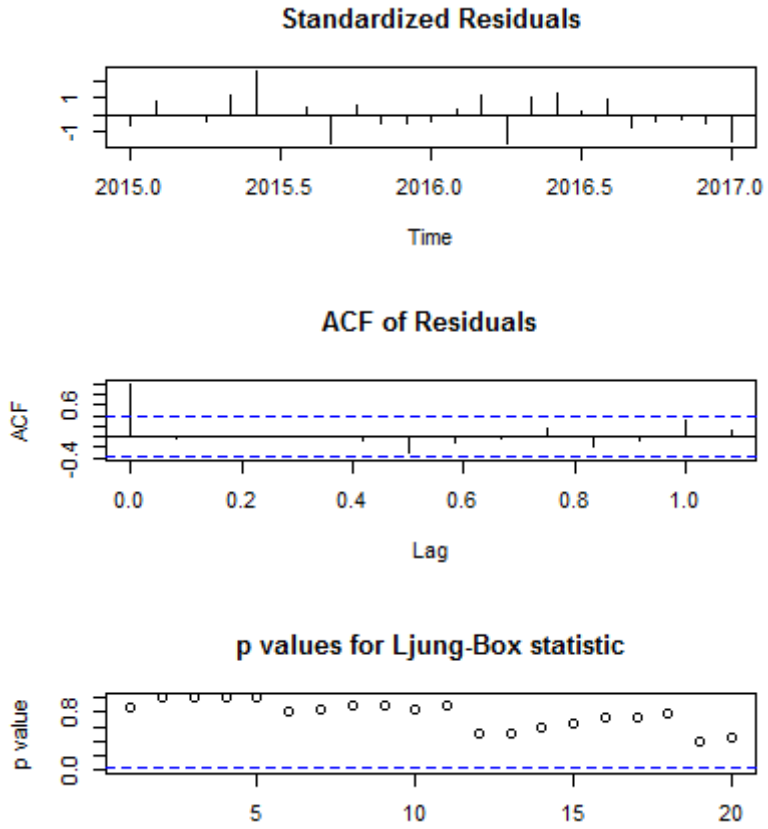

Figura 19. Gràfics d'anàlisi dels residus.

[8] Coefficient(s):

	Estimate	Std. Error	t value	Pr(>|t|)	
ar1	0.9967	0.1849	5.390	7.04e-08	***
ar2	-0.4299	0.1859	-2.313	0.0207	*
intercept	2074.9948	794.3576	2.612	0.0090	**

sigma^2 estimated as 1322396, Conditional Sum-of-Squares = 27770319, AIC = 412.39

L'esperança de la sèrie temporal és $\frac{2074.9948}{1-0.9967+0.4299} = 4789.923$.

En principi això seria suficient però, tenint en compte que hi ha certa recurrència anual, caldria considerar també un model AR(12) i veure si és millor que el AR(2).

El model estimat[9], per tant, seria el següent:

$$X_t = 4744.09 + 0.50X_{t-1} - 0.11\,X_{t-2} - 0.07X_{t-3} - 0.10X_{t-4} - 0.03X_{t-5} - 0.18X_{t-6} + 0.14X_{t-7} - 0.13X_{t-8} + 0.19X_{t-9} - 0.37X_{t-10} + 0.04X_{t-11} + 0.55X_{t-12} + Z_t$$

A la *Figura 20* es pot veure un anàlisi dels residus del model escollit. En principi sembla tant bon ajust com el model anterior. L'única diferència evident entre els dos models és que el model AR(12) té una variància estimada molt inferior a la del model AR(2), per tant, es pot considerar millor.

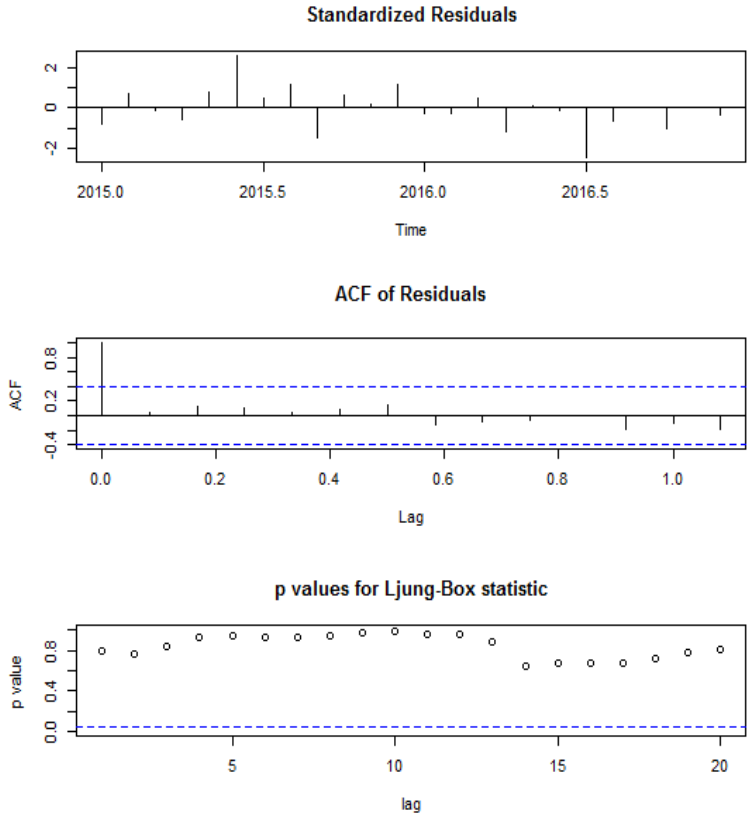

Figura 20. Gràfics d'anàlisi dels residus.

[9] Coefficients:

	ar1	ar2	ar3	ar4	ar5	ar6	ar7	ar8	ar9	ar10	ar11	ar12	intercept
	0.5008	-0.1141	-0.0719	-0.101	-0.0263	-0.1751	0.1415	-0.1290	0.1871	-0.3718	0.0432	0.5471	4744.0895
s.e.	0.1982	0.2466	0.1959	0.200	0.2153	0.1908	0.1873	0.2223	0.2361	0.2971	0.3305	0.2628	184.8174

sigma^2 estimated as 370739: log likelihood = -193.99, aic = 415.98

L'esperança de la sèrie temporal és $\frac{4744.0895}{1-0.5008+0.1141+0.0719+\cdots+0.3718-0.0432-0.5471} = 8330.271$.

Cooker (cuines)

La *Figura 21* mostra la demanda d'aquesta categoria al llarg dels 2 anys. Sembla ser que les vendes han anat disminuint lleugerament durant aquest temps[10], i no s'observa cap patró.

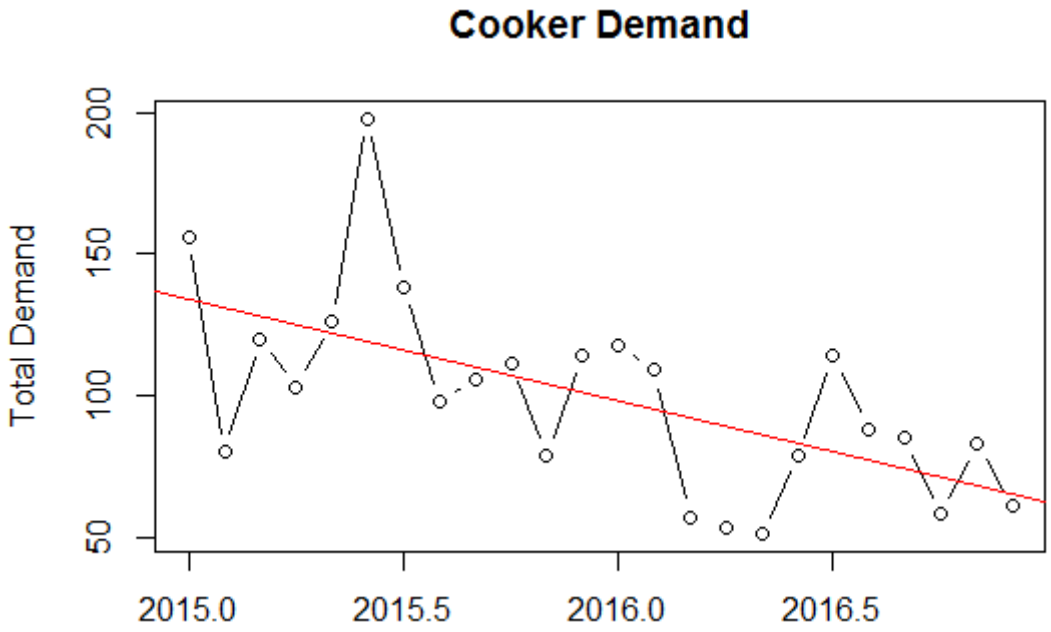

Figura 21. Gràfics de la demanda de Cooker al llarg del temps.

La *Figura 22* mostra la gràfica de les funcions d'auto correlació i d'auto correlació parcial. Pel que s'observa, sembla ser que un model MA(1) seria suficient per ajustar les dades.

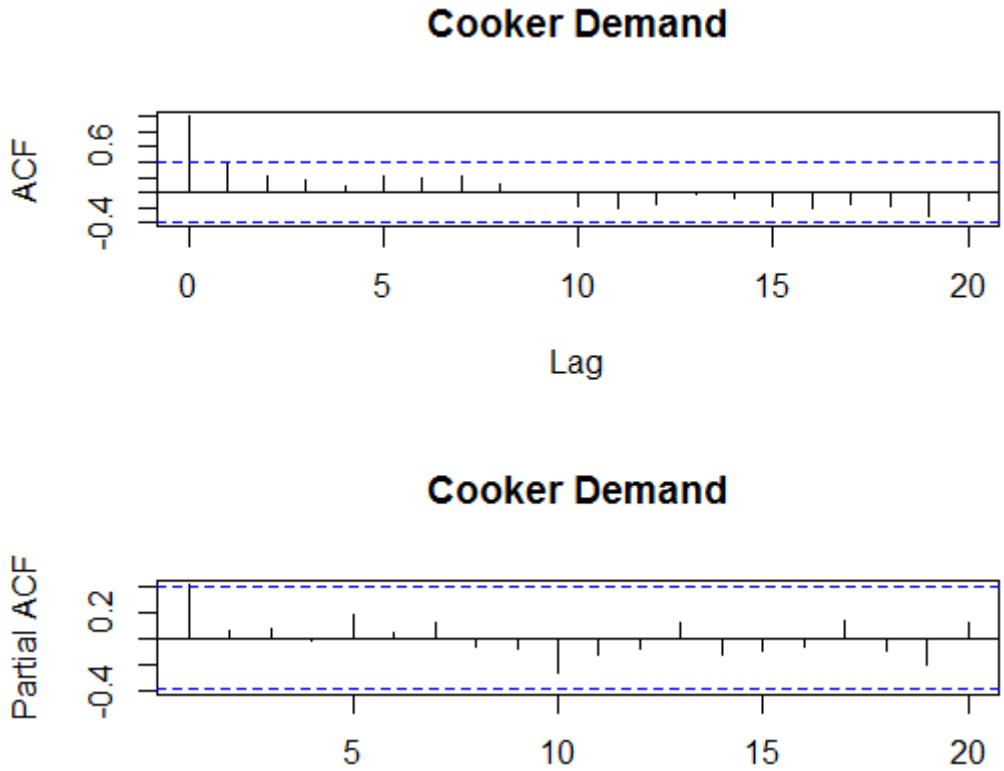

Figura 22. Gràfics d'auto correlació i d'auto correlació parcial de Cooker.

[10]

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	72537.14	20171.71	3.596	0.00161	**
t	-35.93	10.01	-3.591	0.00163	**

Residual standard error: 28.28 on 22 degrees of freedom
Multiple R-squared: 0.3695, Adjusted R-squared: 0.3409
F-statistic: 12.9 on 1 and 22 DF, p-value: 0.001626

Com a comprovació, es pot utilitzar la funció *auto.arima* que calcula l'AIC de tots els possibles models i els compara per donar automàticament el que mostra un millor ajust.

El model estimat[11], per tant, seria el següent:

$$X_t = 94.9258 + Z_t + 0.5201\, Z_{t-1}$$

A la *Figura 23* es pot veure un anàlisi dels residus del model escollit. De la mateixa manera que l'anterior, en el primer gràfic (residus normalitzats) tots es troben dins el rang entre -2 i 2; en el segon gràfic (correlograma dels residus) es veu que té forma de Soroll Blanc; i en el tercer gràfic (p-valors dels retards) els p-valor dels retards estan tots fora del rang de significança. Així doncs, el model és apropiat.

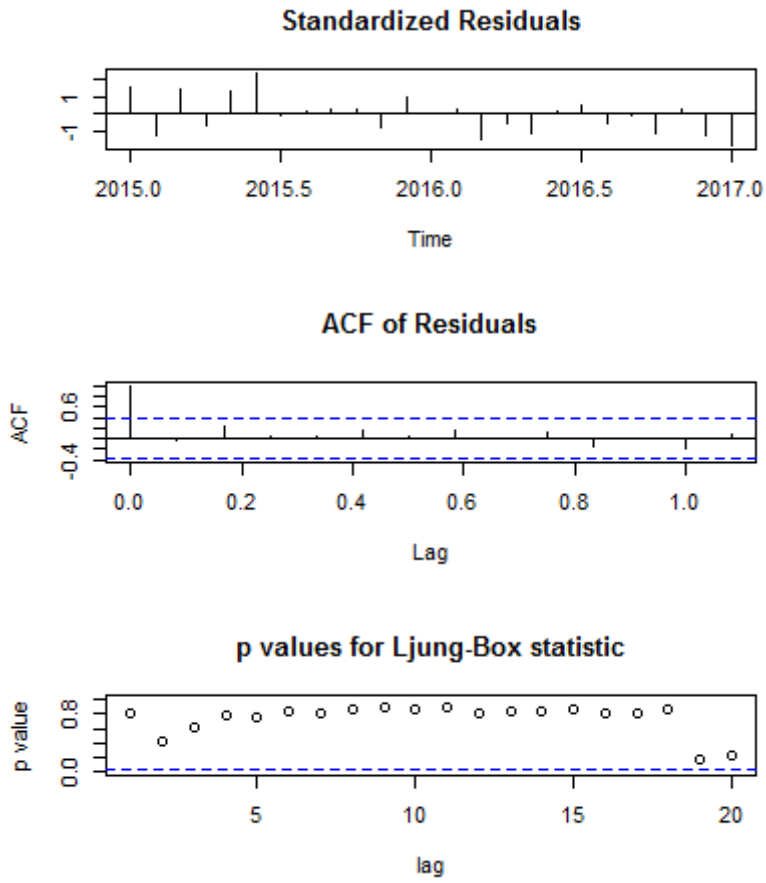

Figura 23. Gràfics d'anàlisi dels residus.

En principi, per fer aquest l'ajust s'hauria de tindre en compte la tendència decreixent comentada anteriorment, però tant traient-la de les dades com utilitzant un model ARIMA provocava que la sèrie temporal es convertís en Soroll Blanc, és a dir, les dades es tornen aleatòries i no es poden modelitzar. Per aquest motiu, he obviat la tendència per així poder ajustar un model amb el qual poder fer les prediccions.

[11] Coefficient(s):

	Estimate	Std. Error	t value	Pr(>|t|)	
ma1	0.5201	0.1682	3.092	0.00199	**
intercept	94.9258	8.6368	10.991	< 2e-16	***

sigma^2 estimated as 833.3, Conditional Sum-of-Squares = 18345.36, AIC = 233.52

L'esperança de la sèrie temporal és 94.9258 .

Dryers (secadores)

La *Figura 24* mostra la demanda d'aquesta categoria al llarg dels 2 anys.

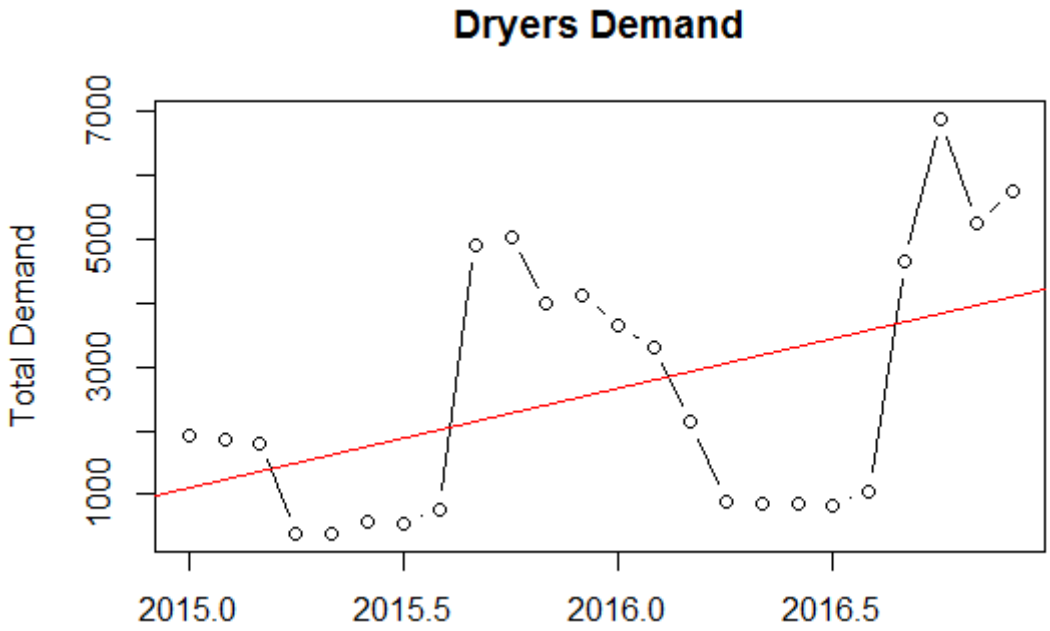

Figura 24. Gràfics de la demanda de Dryers al llarg del temps.

Com era d'esperar, la venda d'assecadores cau en picat des de la primavera fins l'estiu, i remunta als mesos de tardor i hivern, degut a que amb el fred i/o el mal temps és més difícil que la roba es sequi de manera natural, el qual genera una estacionalitat. En general, sembla que aquesta categoria ha tingut un augment en les vendes al llarg d'aquest temps[12]. Podria pensar-se que també hi ha certa recurrència anual, és a dir, al comportament al llarg de l'any es repeteix cada any (tal com mostra la *Figura 25*).

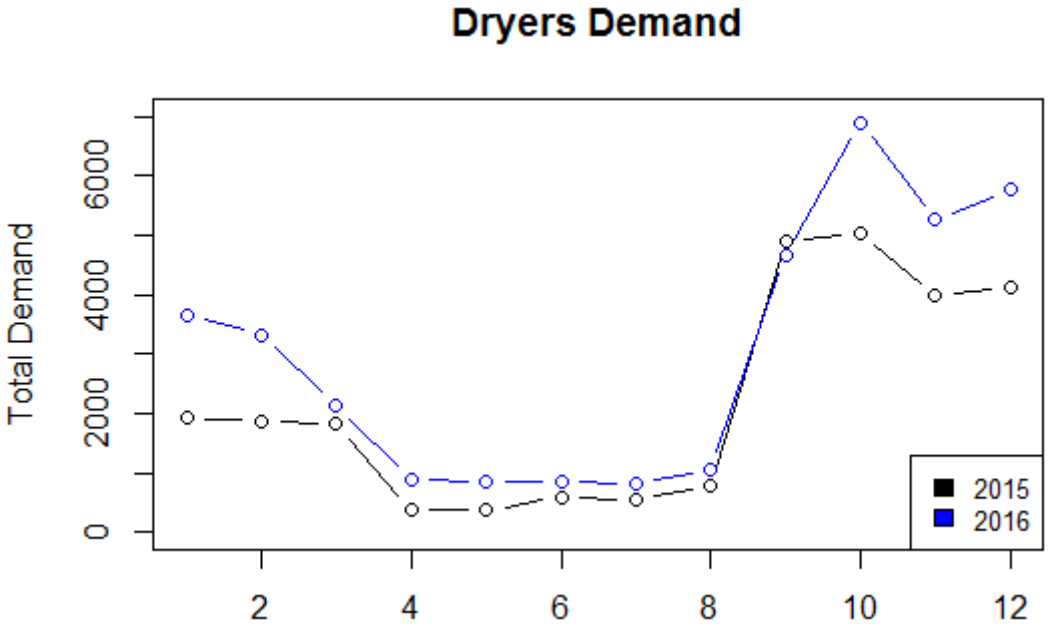

Figura 25. Gràfics de la demanda anual de Dryers.

12

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-3158501.1	1316647.1	-2.399	0.0254	*
t	1568.0	653.1	2.401	0.0252	*

Residual standard error: 1846 on 22 degrees of freedom
Multiple R-squared: 0.2076, Adjusted R-squared: 0.1716
F-statistic: 5.764 on 1 and 22 DF, p-value: 0.02525

La *Figura 26* mostra la gràfica de les funcions d'auto correlació i d'auto correlació parcial. Pel que s'observa, sembla ser que un model ARMA(1,1) seria suficient per ajustar les dades. A més, com a comprovació, es pot utilitzar la funció *auto.arima* que calcula l'AIC de tots els possibles models i els compara per donar automàticament el que mostra un millor ajust.

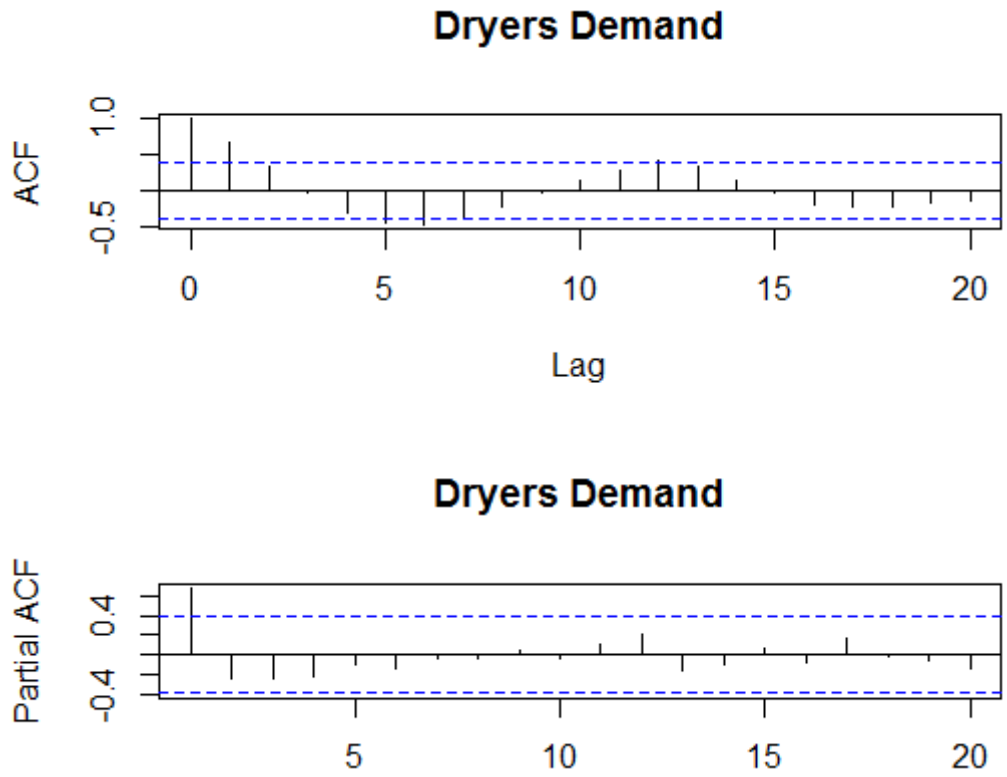

Figura 26. Gràfics d'auto correlació i d'auto correlació parcial de Dryers.

El model estimat[13], per tant, seria el següent:

$$X_t = 1123.7865 + 0.6170X_{t-1} + Z_t + 0.5075\, Z_{t-1}$$

A la *Figura 27* es pot veure un anàlisi dels residus del model escollit. De la mateixa manera que els anteriors, en el primer gràfic (residus normalitzats) tots es troben dins el rang entre -2 i 2; en el segon gràfic (correlograma dels residus) es veu que té forma de Soroll Blanc; i en el tercer gràfic (p-valors dels retards) els p-valor dels retards estan tots fora de rang significança. Així doncs, el model és apropiat.

[13] Coefficient(s):

	Estimate	Std. Error	t value	Pr(>|t|)	
ar1	0.6170	0.2092	2.950	0.00318	**
ma1	0.5075	0.2362	2.149	0.03164	*
intercept	1123.7865	634.1021	1.772	0.07635	.

sigma^2 estimated as 1624482, Conditional Sum-of-Squares = 35746032, AIC = 417.33

L'esperança de la sèrie temporal és $\frac{1123.7865}{1-0.6170} = 2934.168$.

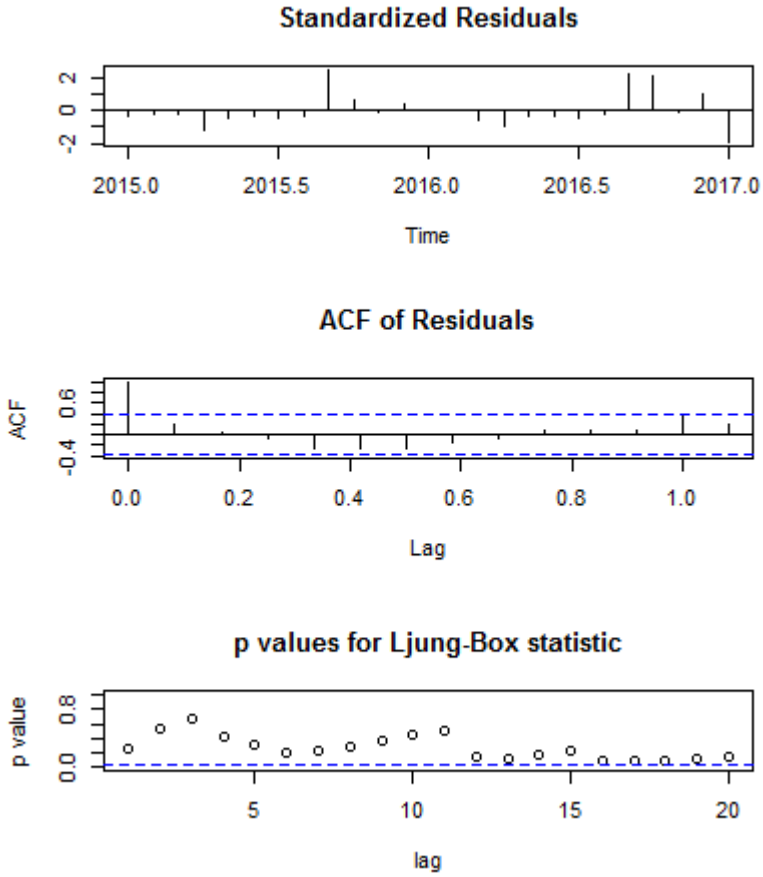

Figura 27. Gràfics d'anàlisi dels residus.

En principi això seria suficient però, tenint en compte que hi ha certa recurrència anual, caldria considerar també un model AR(12) i veure si és millor que el ARMA(1,1).

El model estimat[14], per tant, seria el següent:

$$X_t = 2648.2146 + 0.56X_{t-1} - 0.29\,X_{t-2} + 0.27X_{t-3} - 0.21X_{t-4} - 0.03X_{t-5} - 0.09X_{t-6} + 0.14X_{t-7} - 0.11X_{t-8} + 0.04X_{t-9} - 0.04X_{t-10} - 0.05X_{t-11} + 0.56X_{t-12} + Z_t$$

A la *Figura 28* es pot veure un anàlisi dels residus del model escollit. En principi sembla tant bon ajust com el model anterior. L'única diferència evident entre els dos models és que el model AR(12) té una variància estimada molt inferior a la del model ARMA(1,1), per tant, es pot considerar millor.

[14] Coefficients:

	ar1	ar2	ar3	ar4	ar5	ar6	ar7	ar8	ar9	ar10	ar11	ar12	intercept
	0.5565	-0.2873	0.2695	-0.2142	-0.0305	-0.0856	0.1428	-0.1123	0.0419	-0.0401	-0.0509	0.5578	2648.2146
s.e.	0.1869	0.2246	0.2331	0.2512	0.2628	0.2282	0.2466	0.3023	0.3192	0.3176	0.2854	0.2122	377.1386

sigma^2 estimated as 585873: log likelihood = -198.07, aic = 424.14

L'esperança de la sèrie temporal és $\frac{2648.2146}{1-0.5565+0.2873-0.2695+\cdots+0.0401+0.0509-0.5578} = 10492.13$.

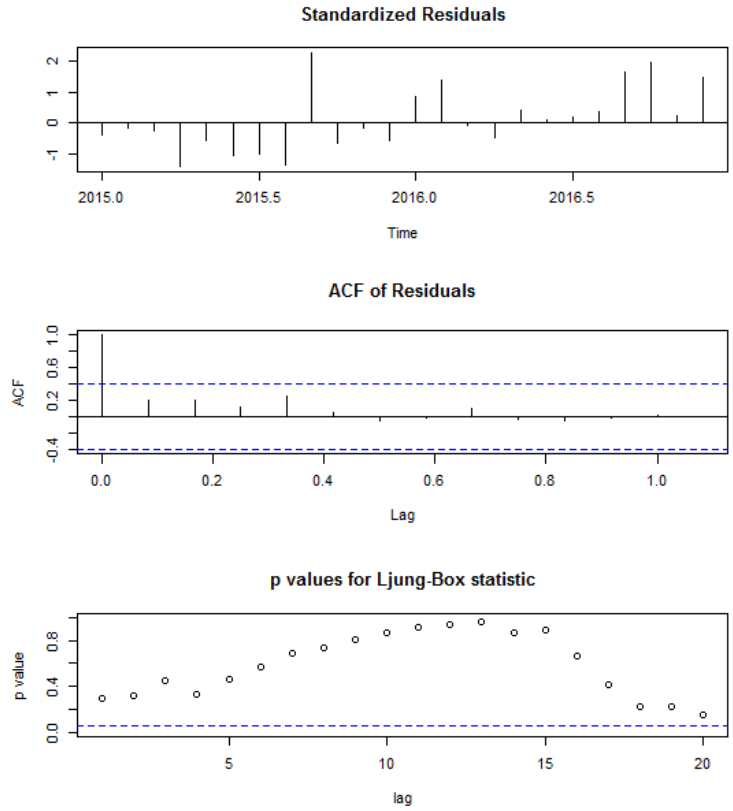

Figura 28. Gràfics d'anàlisi dels residus.

De la mateixa manera que en l'anterior categoria, per fer aquests ajusts s'hauria de tindre en compte la tendència decreixent comentada anteriorment, però tant traient-la de les dades com utilitzant un model ARIMA provocava que la sèrie temporal es convertís en Soroll Blanc, és a dir, les dades es tornen aleatòries i no es poden modelitzar. Per aquest motiu, he obviat la tendència per així poder ajustar un model amb el qual poder fer les prediccions.

Freez (congeladors)

La *Figura 29* mostra la demanda d'aquesta categoria al llarg dels 2 anys. Sembla ser que les vendes han anat disminuint lleugerament durant aquest temps[15], i no s'observa cap patró.

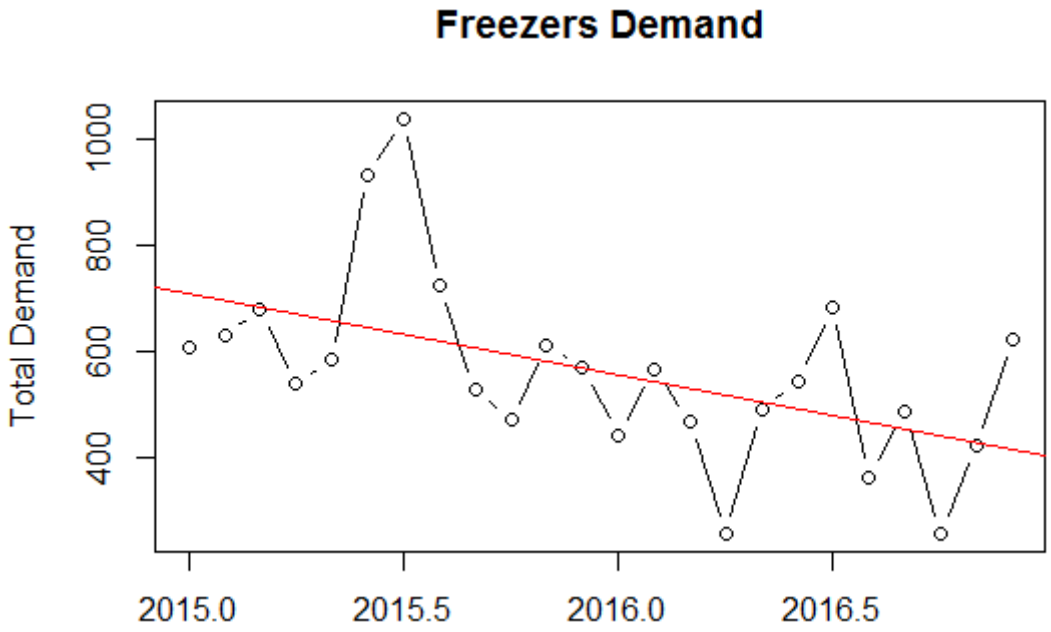

Figura 29. Gràfics de la demanda de Freezers al llarg del temps.

La *Figura 30* mostra la gràfica de les funcions d'auto correlació i d'auto correlació parcial. Pel que s'observa, sembla ser que un model AR(1) seria suficient per ajustar les dades.

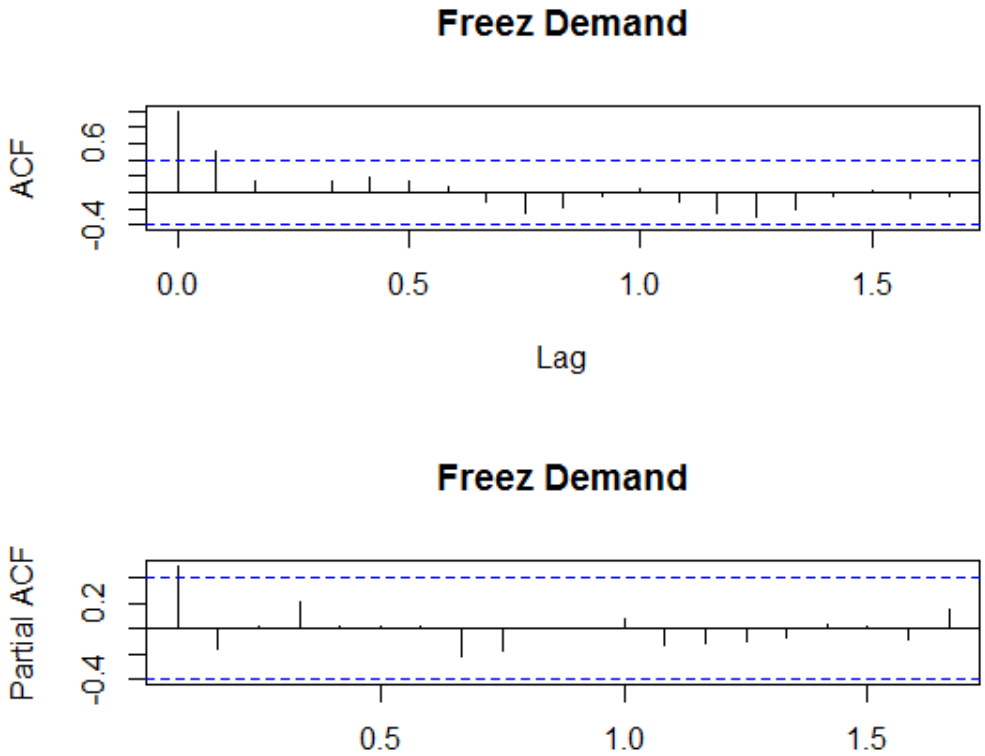

Figura 30. Gràfics d'auto correlació i d'auto correlació parcial de Freezers.

15

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	307679.64	111316.15	2.764	0.0113	*
t	-152.34	55.22	-2.759	0.0115	*

Residual standard error: 156 on 22 degrees of freedom
Multiple R-squared: 0.2571, Adjusted R-squared: 0.2233
F-statistic: 7.612 on 1 and 22 DF, p-value: 0.01145

Com a comprovació, es pot utilitzar la funció *auto.arima* que calcula l'AIC de tots els possibles models i els compara per donar automàticament el que mostra un millor ajust.

El model estimat[16], per tant, seria el següent:

$$X_t = 280.6345 + 0.5006X_{t-1} + Z_t$$

A la *Figura 31* es pot veure un anàlisi dels residus del model escollit. De la mateixa manera que l'anterior, en el primer gràfic (residus normalitzats) tots es troben dins el rang entre -2 i 2; en el segon gràfic (correlograma dels residus) es veu que té forma de Soroll Blanc; i en el tercer gràfic (p-valors dels retards) els p-valor dels retards estan tots fora del rang de significança. Així doncs, el model és apropiat.

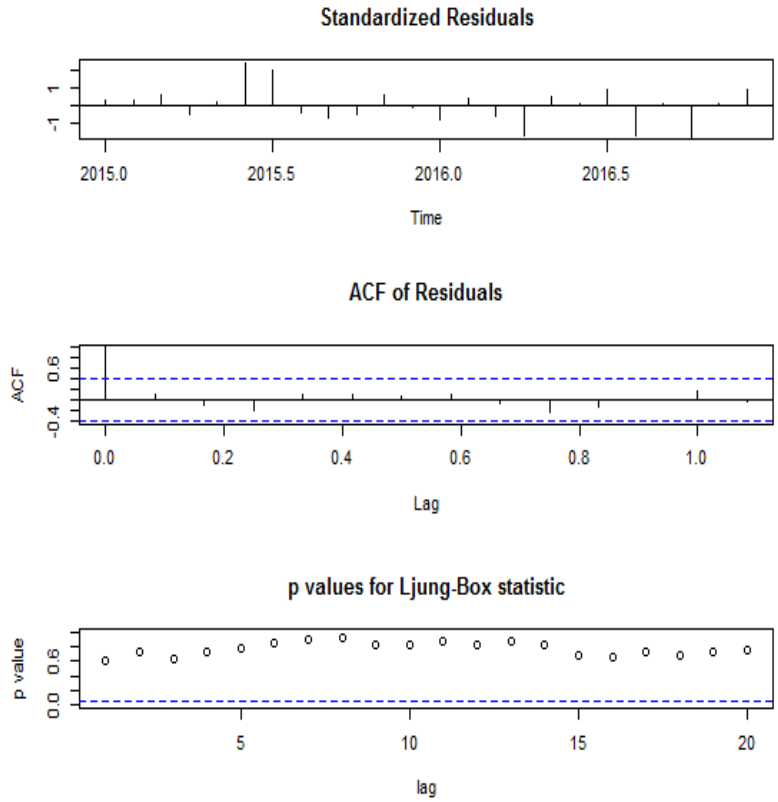

Figura 31. Gràfics d'anàlisi dels residus.

De la mateixa manera que en l'anterior categoria, per fer aquest l'ajust s'hauria de tindre en compte la tendència decreixent comentada anteriorment, però tant traient-la de les dades com utilitzant un model ARIMA provocava que la sèrie temporal es convertís en Soroll Blanc, és a dir, les dades es tornen aleatòries i no es poden modelitzar. Per aquest motiu, he obviat la tendència per així poder ajustar un model amb el qual poder fer les prediccions.

[16] Coefficient(s):

	Estimate	Std. Error	t value	Pr(>|t|)
ar1	0.5006	0.1770	2.829	0.00467 **
intercept	280.6345	104.0300	2.698	0.00698 **

sigma^2 estimated as 24507, Conditional Sum-of-Squares = 539146.3, AIC = 314.67

L'esperança de la sèrie temporal és $\frac{280.6345}{1-0.5006} = 561.9433$.

Refr (neveres)

La *Figura 32* mostra la demanda d'aquesta categoria al llarg dels 2 anys. De la mateixa manera que amb els Combi, sembla ser que les vendes augmenten als mesos d'estiu i disminueixen a l'hivern, generant estacionalitat en la sèrie. També sembla que, en general, aquesta categoria ha tingut un descens en les vendes al llarg d'aquest temps, tot i que no és significativament diferent de zero[17]. Podria pensar-se que, a més, hi ha certa recurrència anual, és a dir, al comportament al llarg de l'any es repeteix cada any (tal com mostra la *Figura 33*).

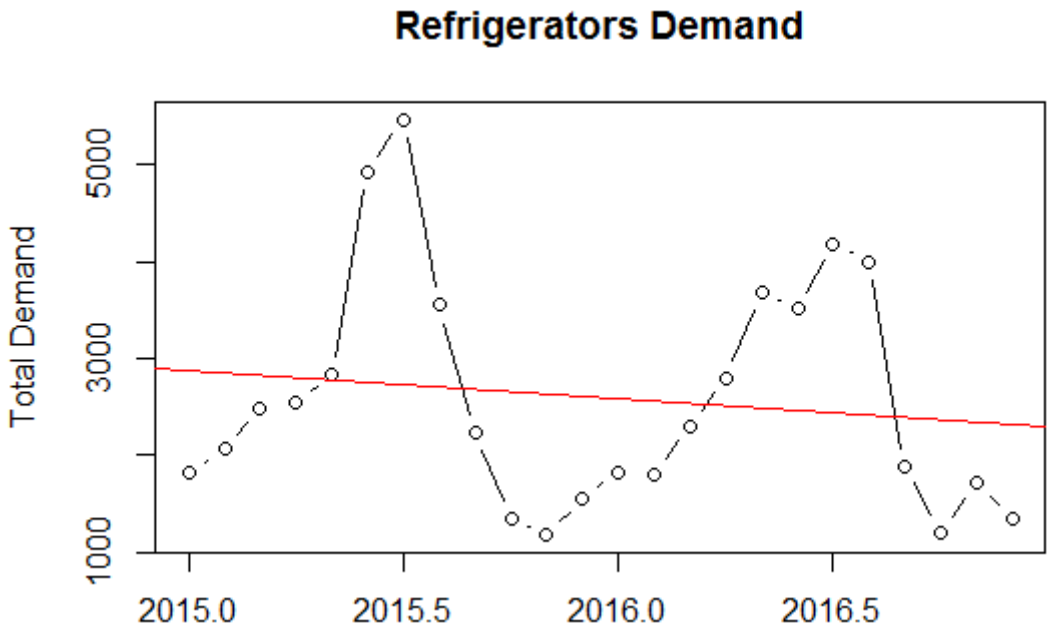

Figura 32. Gràfics de la demanda de Refr al llarg del temps.

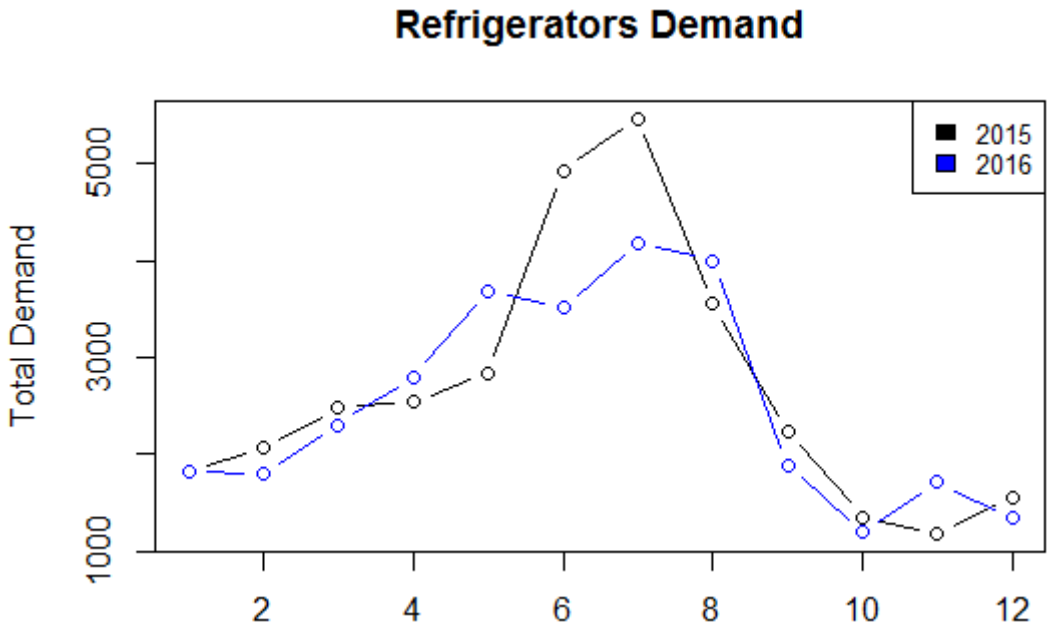

Figura 33. Gràfics de la demanda de Refr al llarg del temps.

[17]

	Estimate	Std. Error	t value	Pr(>|t|)	Residual standard error: 1214 on 22 degrees of freedom
(Intercept)	589099.6	865949.9	0.680	0.503	Multiple R-squared: 0.02043, Adjusted R-squared: -0.0241
t	-290.9	429.5	-0.677	0.505	F-statistic: 0.4587 on 1 and 22 DF, p-value: 0.5053

La *Figura 34* mostra la gràfica de les funcions d'auto correlació i d'auto correlació parcial. Pel que s'observa, sembla ser que un model ARMA(1,1) seria suficient per ajustar les dades. Com a comprovació, es pot utilitzar la funció *auto.arima* que calcula l'AIC de tots els possibles models i els compara per donar automàticament el que mostra un millor ajust.

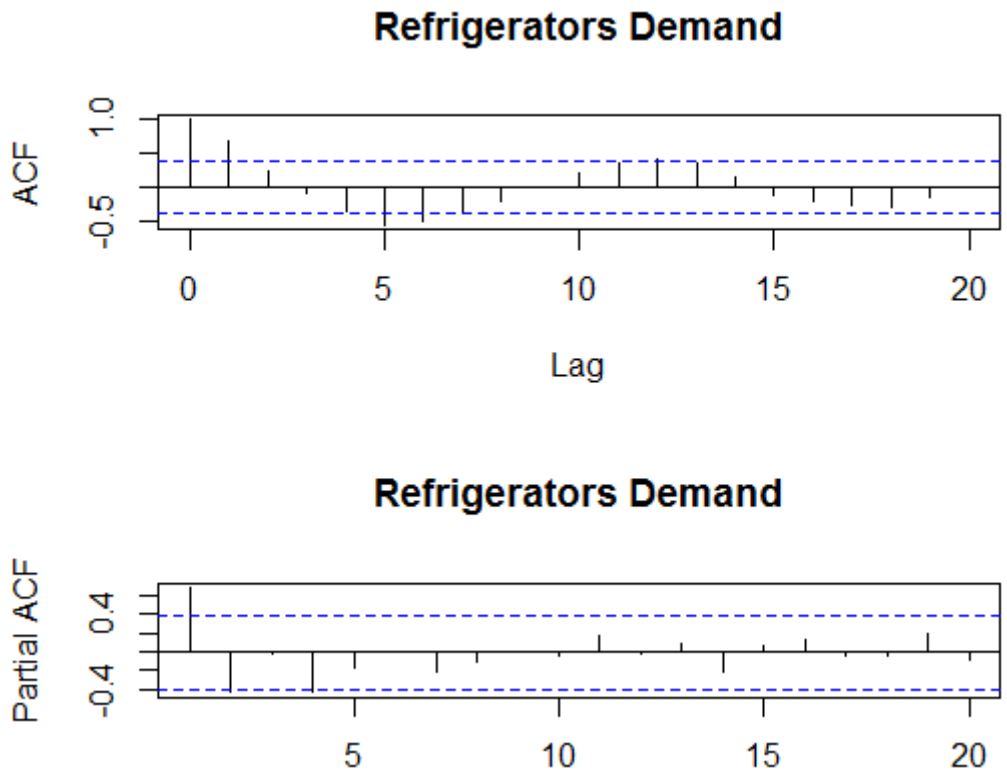

Figura 34. Gràfics d'auto correlació i d'auto correlació parcial de Refr.

El model estimat[18], per tant, seria el següent:

$$X_t = 1175.0191 + 0.5337X_{t-1} + Z_t + 0.7679Z_{t-1}$$

A la *Figura 35* es pot veure un anàlisi dels residus del model escollit. De la mateixa manera que l'anterior, en el primer gràfic (residus normalitzats) tots es troben dins el rang entre -2 i 2; en el segon gràfic (correlograma dels residus) es veu que té forma de Soroll Blanc; i en el tercer gràfic (p-valors dels retards) els p-valor dels retards estan tots fora del rang de significança. Així doncs, el model és apropiat.

[18] Coefficient(s):

	Estimate	Std. Error	t value	Pr(>|t|)	
ar1	0.5337	0.1843	2.896	0.00378	**
ma1	0.7679	0.1184	6.486	8.84e-11	***
intercept	1175.0191	522.7736	2.248	0.02460	*

sigma^2 estimated as 472981, Conditional Sum-of-Squares = 10406844, AIC = 387.71

L'esperança de la sèrie temporal és $\frac{1175.0191}{1-0.5337} = 2519.878$.

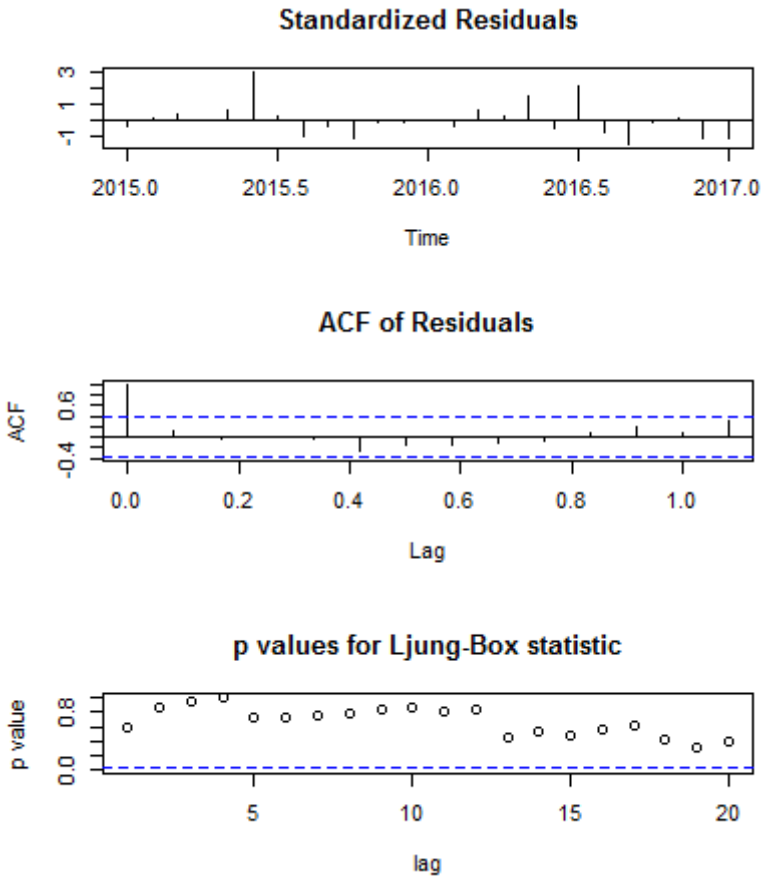

Figura 35. Gràfics d'anàlisi dels residus.

En principi això seria suficient però, tenint en compte que hi ha certa recurrència anual, caldria considerar també un model AR(12) i veure si és millor que el ARMA(1,1).

El model estimat[19], per tant, seria el següent:

$$X_t = 2591.19 + 0.67X_{t-1} - 1.00\,X_{t-2} + 0.51X_{t-3} - 0.74X_{t-4} + 0.27X_{t-5} - 0.15X_{t-6} - 0.56X_{t-7} + 0.32X_{t-8} - 0.71X_{t-9} + 0.61X_{t-10} - 0.60X_{t-11} + 0.54X_{t-12} + Z_t$$

A la *Figura 36* es pot veure un anàlisi dels residus del model escollit. En principi sembla tant bon ajust com el model anterior. L'única diferència evident entre els dos models és que el model AR(12) té una variància estimada molt inferior a la del model ARMA(1,1), per tant, es pot considerar millor.

[19] Coefficients:

	ar1	ar2	ar3	ar4	ar5	ar6	ar7	ar8	ar9	ar10	ar11	ar12	intercept
	0.6654	-1.0010	0.5082	-0.7431	0.2742	-0.1446	-0.5579	0.3187	-0.7055	0.6073	-0.5970	0.5420	2591.1885
s.e.	0.1902	0.1835	0.2536	0.2199	0.2473	0.1810	0.2694	0.2891	0.2290	0.2696	0.2043	0.1999	40.2138

sigma^2 estimated as 96488: log likelihood = -178.91, aic = 385.83

L'esperança de la sèrie temporal és $\frac{2591.1885}{1-0.6654+1.001-0.5082+\cdots-0.6073+0.597-0.542} = 1413.401$.

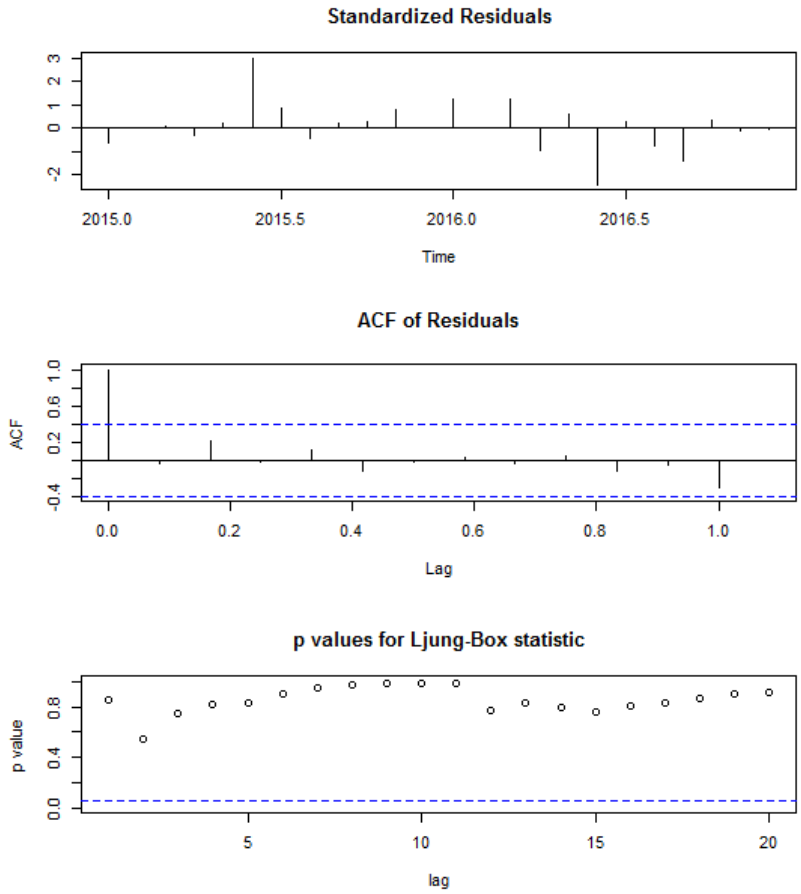

Figura 36. Gràfics d'anàlisi dels residus.

Promoció i Trimestre

Més enllà dels retards en la demanda, també hi ha les variables promoció (si en el mes té o no promoció) i trimestre en el qual es troba. Per veure si són significatives dins el model on la demanda és la dependent i la categoria, la promoció i el trimestre són les variables explicatives cal fer una ANOVA.

Una ANOVA és una taula on es comparen diferents nivells d'una variable categòrica, per veure si la seva influència en la variable resposta és diferent segons el nivell o si tots es poden considerar iguals i, per tant, estar en un nivell o un altre produeix indistintament el mateix valor de la variable resposta. Els p-valors que produeix són fruit d'un contrast que té com a hipòtesi nul·la que tots els nivells de la variable categòrica es poden considerar iguals i com a hipòtesi alternativa que com a mínim un dels nivells és diferent de la resta. A més, aquesta taula funciona com un model lineal on es busquen les significativitats dels coeficients, per tant, el fet d'incloure només una variable categòrica o varies, pot canviar el resultat dels p-valors.

Les variables categòriques que interessa veure si els seus nivells són diferents entre ells són: categoria (32 nivells), promoció (2 nivells: si hi ha o no) i trimestre (4 nivells).

Fent ANOVA de manera individual, la categoria y la promoció surten significatives mentre que el trimestre no:

```
Response: Total
            Df       Sum Sq     Mean Sq   F value      Pr(>F)
Category    31  7642678051  246538002   410.07  < 2.2e-16 ***
Residuals 736   442488594       601207
```

```
Response: Total
             Df       Sum Sq       Mean Sq   F value      Pr(>F)
Prom          1  1537968547  1537968547   179.94  < 2.2e-16 ***
Residuals  766  6547198098       8547256
```

```
Response: Total
             Df       Sum Sq   Mean Sq   F value   Pr(>F)
Trim          3      7313635   2437878   0.2306  0.8751
Residuals 764  8077853010  10573106
```

Fent ANOVA de manera aparellada, la categoria y la promoció surten sempre significatives mentre que el trimestre només ho és juntament amb la categoria:

```
Response: Total
            Df       Sum Sq     Mean Sq   F value       Pr(>F)
Category    31  7642678051  246538002  441.938   < 2.2e-16 ***
Prom         1    32464323   32464323   58.195  7.366e-14 ***
Residuals  735   410024272     557856
```

```
Response: Total
             Df       Sum Sq     Mean Sq     F value       Pr(>F)
Category     31  7642678051  246538002   415.2637  < 2.2e-16 ***
Trim          3     7313635     2437878     4.1063   0.006651 **
Residuals   733   435174959      593690
```

```
Response: Total
             Df       Sum Sq       Mean Sq     F value   Pr(>F)
Prom          1  1537968547  1537968547  179.9952  <2e-16 ***
Trim          3    27747744     9249248     1.0825  0.3557
Residuals   763  6519450354     8544496
```

Fent ANOVA de manera conjunta, totes les variables surten significatives:

```
Response: Total
             Df       Sum Sq     Mean Sq    F value        Pr(>F)
Category     31 7642678051  246538002  450.7677   < 2.2e-16 ***
Prom          1   32464323   32464323   59.3575   4.282e-14 ***
Trim          3    9672128    3224043    5.8948  0.0005613 ***
Residuals   732  400352143     546929
```

Així doncs, de manera global, el trimestre es torna útil quan es mira amb la categoria o amb la categoria y la promoció, però no per separat.

Fent un test de significança de la variable promoció, per categoria, és significativa al 95% de confiança per les següents categories, amb els seus respectius p-valors: *Dish* – 0.00013, *Dryers* – 0.03823, *Hob* – 0.04339, *WashDry* – 0.01943 i *Washer* – 0.00801 .

La *Figura 37* mostra per cada categoria amb la promoció significativa, els seus gràfics de caixa de la demanda segons si aquell mes hi havia o no promoció. Sembla que, en general, la variabilitat en les vendes dels mesos que no hi ha promoció és més elevada que la dels mesos en que hi ha promoció.

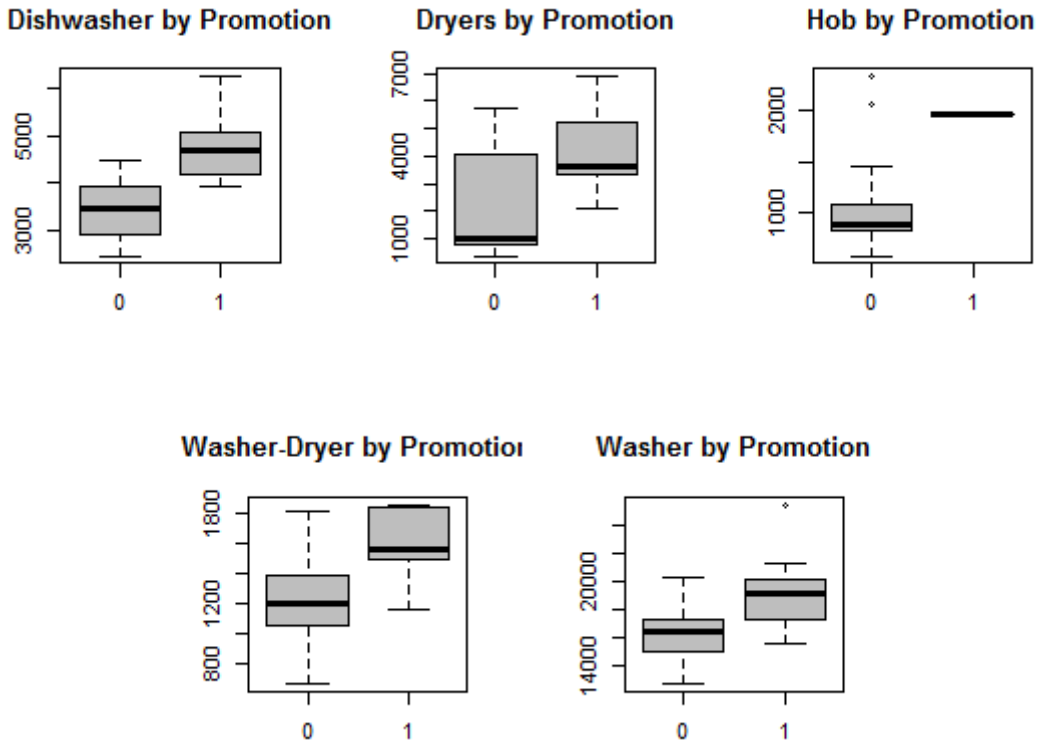

Figura 37. Gràfics de caixa de la demanda segons si hi ha o no promoció.

Aquesta informació sobre les promocions implica que, per aquestes categories, el fet d'oferir promocions realment produeix un augment en les vendes i, a més, en redueix la variabilitat. Seria interessant tindre-ho en compte de cara a incentivar promocions en aquestes cinc categories, prioritzant-les sobre la resta de productes.

Fent un test de significança de la variable trimestre, per categoria, és significativa al 95% de confiança per les següents categories, amb els seus respectius p-valors: *Combi* – 0.00000021, *Dryers* – 0.00000024 , *Refr* – 0.00000005 i *WashDry* – 0.00590695 .

La *Figura 38* mostra per cada categoria amb trimestres significatius, els seus gràfics de caixa de la demanda segons cada trimestre. L'explicació de per què el trimestre és significatiu en aquestes categories és la mateixa que el motiu de que mostrin correlació amb els seus retards.

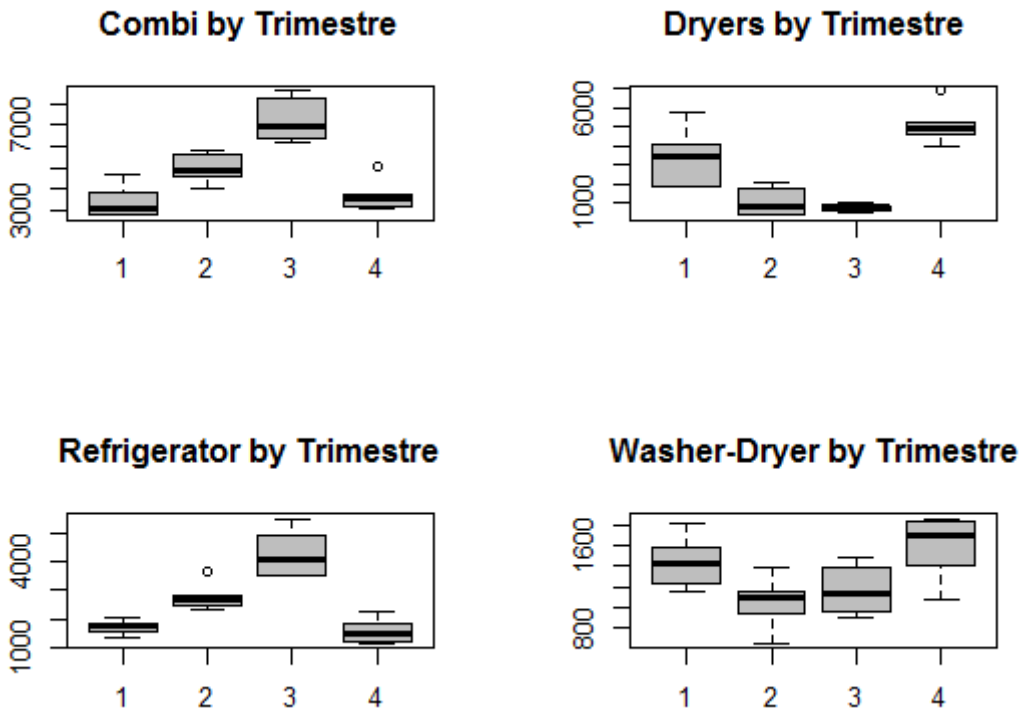

Figura 38. Gràfics de caixa de la demanda segons el trimestre.

La informació que se'n pot extreure és que són categories estacionals, és a dir, es venen més en certs moments de l'any. Es podria considerar aprofitar la distribució de les vendes per incentivar encara més els trimestres de major demanda per elevar-la i reduir-ne la variabilitat.

Prediccions

Donat que el model general era equivalent a fer una predicció aleatòria, només es poden calcular estimacions per aquelles categories que han mostrat una mínima significança. Per la resta es sobreentén que la millor predicció possible es l'esperança de la categoria, és a dir, la mitjana de vendes produïdes fins el moment actual.

Una vegada obtinguts tots els models d'auto correlació adients per cadascuna de les categories significatives, les prediccions que es produeixen pels 12 mesos següents són:

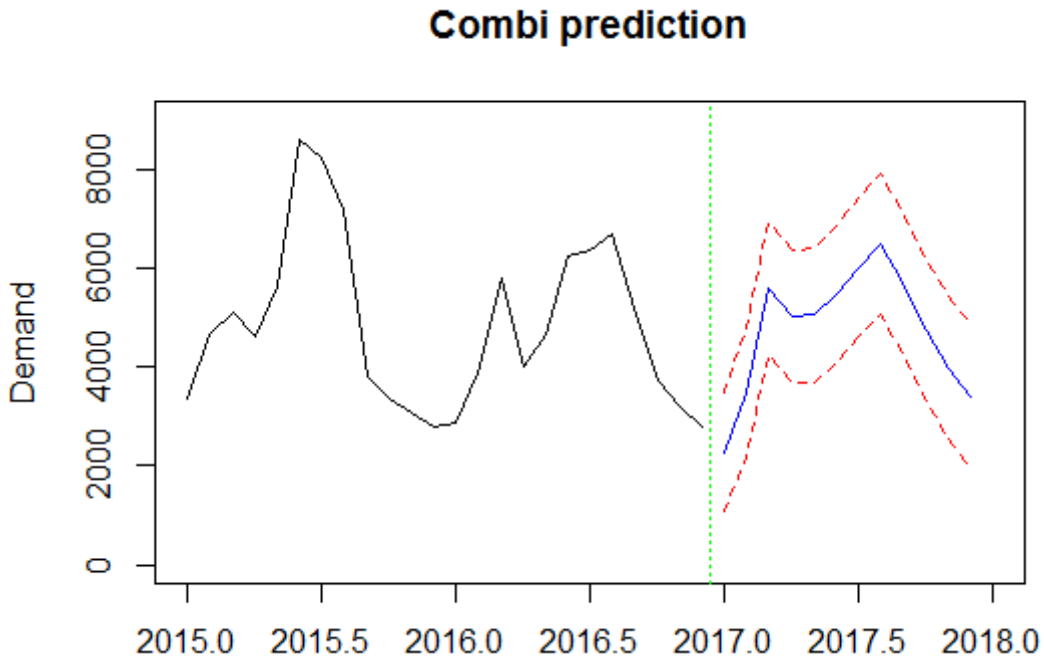

Figura 39. Predicció de la demanda de Combi pel 2017.

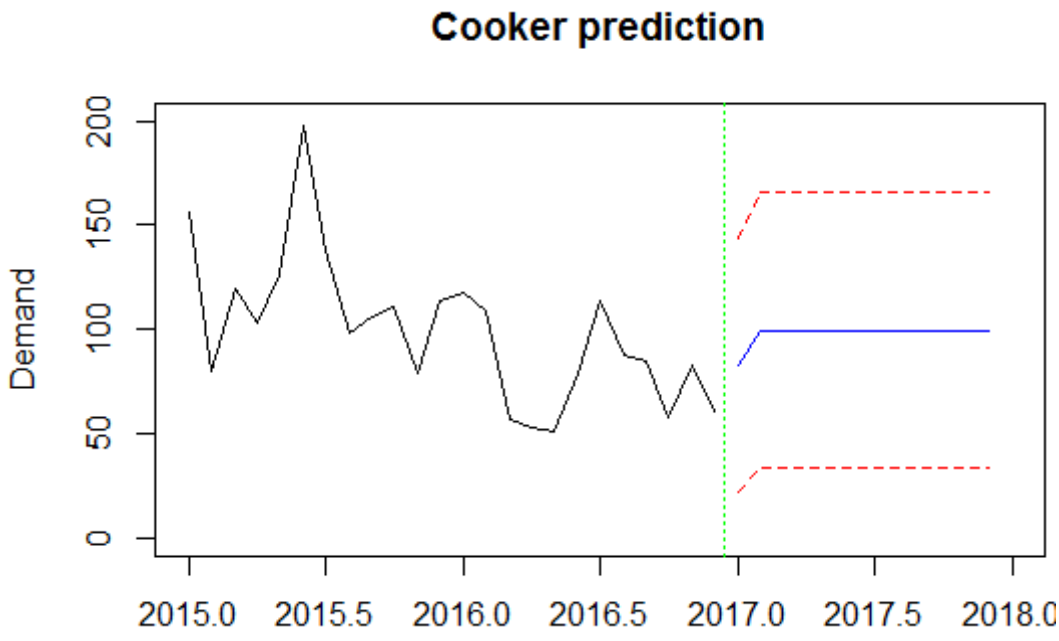

Figura 40. Predicció de la demanda de Cooker pel 2017.

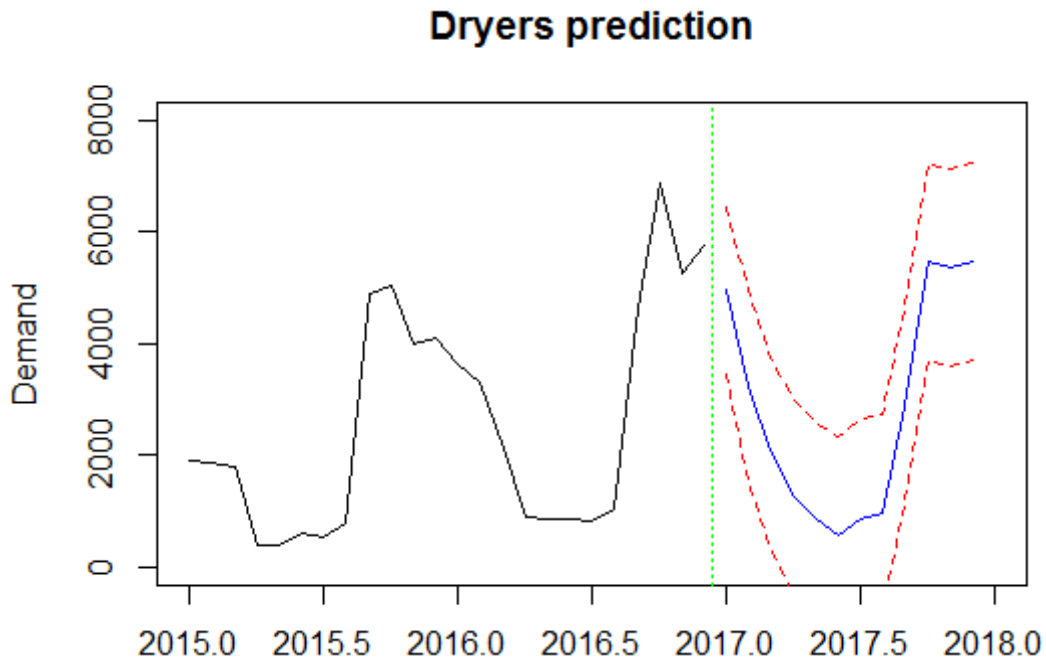

Figura 41. Predicció de la demanda de Dryers pel 2017.

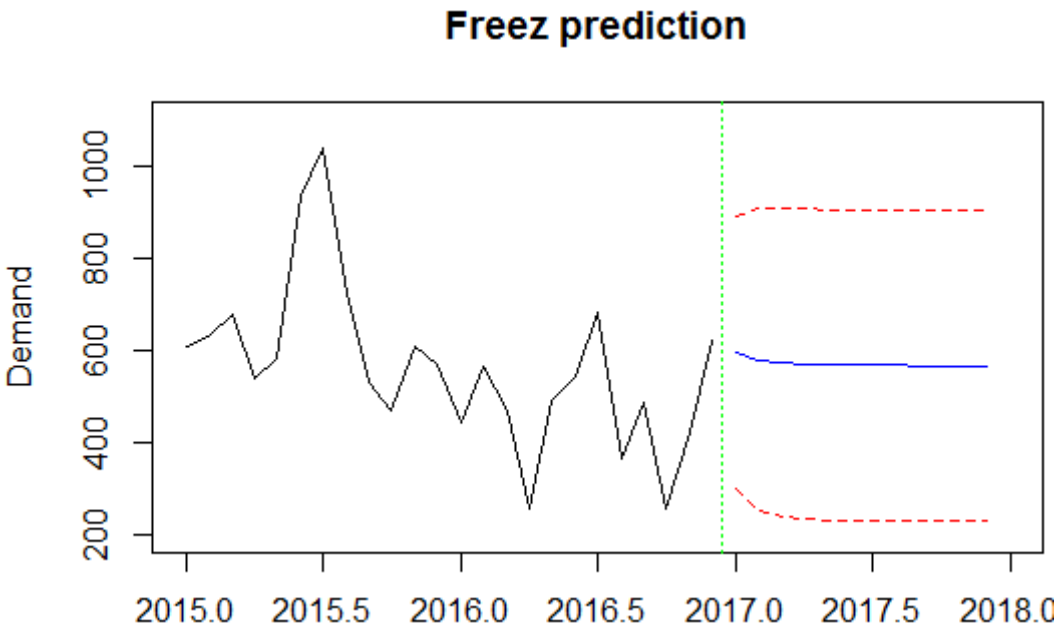

Figura 42. Predicció de la demanda de Freezers pel 2017.

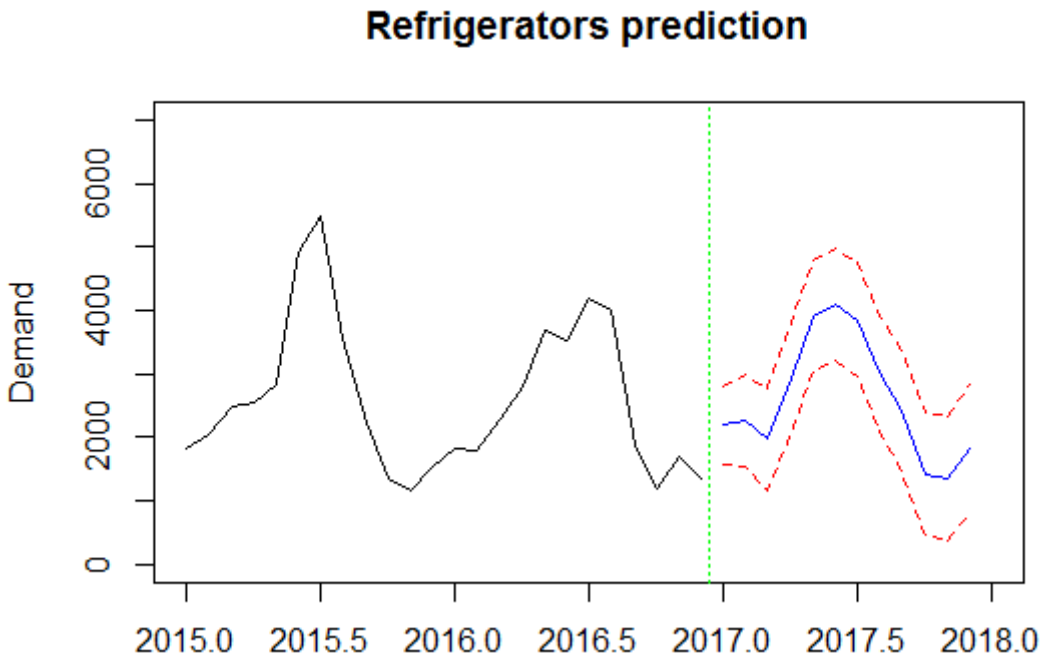

Figura 43. Predicció de la demanda de Refrigerators pel 2017.

Conclusions

Com es pot observar, cal diferenciar entre les sèries que mostren anualitat i les que no. Les primeres mostren un ajust molt més acurat en relació a les dades dels anys anteriors mentre que les segones tenen una predicció pràcticament sense forma. Això es degut a que, al no tindre una forma anual definida, el valor predit tendeix a la mitjana de la sèrie.

Així doncs, mentre que per les categories amb anualitat es pot considerar que serà prou fiable en relació al que passarà realment, en les categories més indefinides, i tenint en compte l'amplitud de les bandes de confiança, les estimacions s'haurien de prendre més com una referència aproximada del que pot passar que com a prediccions puntuals.

Bibliografia

Time Series Analisys and Its Aplications: With R Examples, Second Edition (*Robert H. Shumway, David S. Stoffer*)

Time Series: Theory and Methods, Second Edition (*Peter J. Brockwell, Richard A. Davis*)

Methods of Multivariate Analysis, Third Edition (*Alvin C. Rencher, William F. Chistensen*)

ANNEX

Codi R

```
> dades<-read.table("datos.txt",header=T,sep="")
> dades2<-
data.frame(Category=as.factor(dades$Category),Month=as.factor(dades$Month),
+ Demand=dades$Demand,Promotion=dades$Promotion,Prom=as.factor(dades$Prom))
> dades2$Trim<-as.factor(c(rep(1,2*32),rep(2,3*32),rep(3,3*32),rep(4,3*32),
+ rep(1,3*32),rep(2,3*32),rep(3,3*32),rep(4,3*32),rep(1,2*32)))
> dades2$Total<-dades2$Demand+dades2$Promotion
> dades2<-subset(dades2,Month!="1701")
> names(dades2)
> summary(dades2)
>
> # Dades sense categoria
> dadesT<-read.table("datosTotal.txt",header=T,sep="")
> dadesT$Total<-dadesT$Demand+dadesT$Promotion
> dadesT<-subset(dadesT,Month!="1701")
> dadesT2<-ts(dadesT[,-1],start=c(2015,1),frequency=12)
> dadesT2
> plot(dadesT2[1:12,4],type="b",xlab="Month",ylab="Total Demand",
main="Monthly Demand",ylim= c(30000,55000))
> lines(dadesT2[13:24,4],type="b",col="blue")
> legend("topleft",legend=c("2015","2016"),fill=c("black","blue"),cex=0.8)
>
> # Tendencia i St
>
> meanT2<-numeric()
> for(i in 1:12){
>   meanT2[i]<-(dadesT2[12+i,"Total"]+dadesT2[i,"Total"])/2
> }
> newT2<-dadesT2[,"Total"]-rep(meanT2,2)
> trend<-summary(lm(newT2~t))$coefficients[2,1]
> pureT2<-newT2-trend
> plot(pureT2,main="Total (arranged)",ylab="")
>
> par(mfrow=c(2,1))
> acf(pureT2,main="Series Total (arranged)")
> pacf(pureT2,main="Series Total (arranged)")
```

```
> layout(1)
> Box.test(pureT2,lag=12)
>
> # Independencia
> Box.test(dadesT2[,4]) # H0: independencia Demanda
> par(mfrow=c(2,1)) # Demanda
> acf(dadesT2[,4],20,main="Total Demand") # autocorrelacio
> pacf(dadesT2[,4],20,main="Total Demand") # autocorrelacio parcial
> layout(1)
>
> # Model 3 mesos anteriors
> t<-1:nrow(dadesT2)
> t1<-t[1:(length(t)-1)]
> t2<-t[1:(length(t)-2)]
> t3<-t[1:(length(t)-3)]
> dadesT3<-
cbind(Demand=dadesT2[,4],Promo=dadesT2[,3],Dem1=c(0,dadesT2[t1,4]),
+ Dem2=c(0,0,dadesT2[t2,4]),Dem3=c(0,0,0,dadesT2[t3,4]))
> mod<-lm(Demand~Dem1+Dem2+Dem3,data=dadesT3)
> summary(mod)
>
> # Grafic demanda vs categoria global
> plot(dades2$Category,dades2$Total,col="grey",main="Demand x Category")
> dades2Cat<-subset(dades2,dades2$Total<10000)
> plot(dades2Cat$Category,dades2Cat$Total,col="grey",main="Demand x
Category")
>
> # Grafic demanda vs categoria
> nCat<-length(levels(dades2$Category))
> for(i in 1:nCat){
+ layout(matrix(c(1,1,2,3),ncol=2,byrow=T))
+ catname<-levels(dades$Category)[i]
+ plot(as.numeric(dades2$Month[dades2$Category==catname]),dades2$Total
[dades2$Category==catname],abline(h=mean(dades2$Total[dades2$Category==
catname]),col="red")
+ hist(dades2$Total[dades2$Category==catname],freq=F,main=catname,xlab=
"Demand")
+ boxplot(dades2$Total[dades2$Category==catname]~dades2$Trim[dades2$Category
==catname],main=catname)
+ layout(1)
+ }
```

```
>
> # Model per categoria
> summary(lm(Total~Prom+Category+Month,data=dades2))
>
> # ANOVA Prom vs no Prom
> anova(lm(Total~Category,data=dades2))
> anova(lm(Total~Prom,data=dades2))
> anova(lm(Total~Trim,data=dades2))
> anova(lm(Total~Category+Prom,data=dades2))
> anova(lm(Total~Category+Trim,data=dades2))
> anova(lm(Total~Prom+Trim,data=dades2))
> anova(lm(Total~Category+Prom+Trim,data=dades2))
> anProm<-numeric(0)
> anTrim<-numeric(0)
> for(i in 1:nCat){
+ catname<-levels(dades$Category)[i]
+ if(any(dades2$Prom[dades2$Category==catname]==1)){
+ anProm<-
c(anProm,anova(lm(Total~Prom,data=subset(dades2,Category==catname)))$"Pr(>F)"
[1])
+ anTrim<-
c(anTrim,anova(lm(Total~Trim,data=subset(dades2,Category==catname)))$"Pr(>F)"
[1])
+ }
+ if(!any(dades2$Prom[dades2$Category==catname]==1)){
+ anProm<-c(anProm,1)
+ anTrim<-c(anTrim,1)
+ }
+ }
> names(anProm)<-levels(dades2$Category)
> names(anTrim)<-levels(dades2$Category)
> anProm[anProm<0.05]
> anTrim[anTrim<0.05]
> layout(matrix(c(1,1,2,2,3,3,6,4,4,5,5,7),byrow=T,nrow=2))
> layout(matrix(c(1,1,2,2,3,3,6,4,4,5,5,7),byrow=T,nrow=2))
> plot(subset(dades2,Category=="Dish")$Prom,subset(dades2,Category==
"Dish")$Total,col="grey",main="Dishwasher by Promotion")
> plot(subset(dades2,Category=="Dryers")$Prom,subset(dades2,Category==
"Dryers")$Total,col="grey",main="Dryers by Promotion")
> plot(subset(dades2,Category=="Hob")$Prom,subset(dades2,Category==
"Hob")$Total,col="grey",main="Hob by Promotion")
```

```
> plot(subset(dades2,Category=="WashDry")$Prom,subset(dades2,Category==
"WashDry")$Total,col="grey",main="Washer-Dryer by Promotion")
> plot(subset(dades2,Category=="Washer")$Prom,subset(dades2,Category==
"Washer")$Total,col="grey",main="Washer by Promotion")
> layout(1)
> par(mfrow=c(2,2))
> plot(factor(subset(dades2,Category=="Combi")$Trim),subset(dades2,Category==
"Combi")$Total,col="grey",main="Combi by Trimestre")
> plot(factor(subset(dades2,Category=="Dryers")$Trim),subset(dades2,Category
=="Dryers")$Total,col="grey",main="Dryers by Trimestre")
> plot(factor(subset(dades2,Category=="Refr")$Trim),subset(dades2,Category==
"Refr")$Total,col="grey",main="Refrigerator by Trimestre")
> plot(factor(subset(dades2,Category=="WashDry")$Trim),subset(dades2,Category
=="WashDry")$Total,col="grey",main="Washer-Dryer by Trimestre")
> layout(1)
>
> # Independencia
> ind<-numeric(0)
> for(i in 1:nCat){
+ catname<-levels(dades$Category)[i]
+ ind<-c(ind,Box.test(dades2$Total[dades2$Category==catname])$p.value)
+ }
> names(ind)<-levels(dades2$Category)
> ind<-na.omit(ind)
> ind[ind<0.05] # p-valor categories no independents (H0: independ_encia)
> par(mfrow=c(2,3))
> boxplot(subset(dades2,Category=="Combi")$Total,col="grey",main="Demand",
xlab="Combi")
> boxplot(subset(dades2,Category=="Cooker")$Total,col="grey",main="Demand",
xlab="Cooker")
> boxplot(subset(dades2,Category=="Dryers")$Total,col="grey",main="Demand",
xlab="Dryers")
> boxplot(subset(dades2,Category=="Freez")$Total,col="grey",main="Demand",
xlab="Freezers")
> boxplot(subset(dades2,Category=="Refr")$Total,col="grey",main="Demand",
xlab="Refrigerators")
> layout(1)
> par(mfrow=c(2,1)) # Demanda Other (BIFreez)
> acf(dades2$Total[dades2$Category=="BIFreez"],20,main="BI Freezers Demand")
> pacf(dades2$Total[dades2$Category=="BIFreez"],20,main="BI Freezers Demand")
> layout(1)
> Box.test(dades2$Total[dades2$Category=="BIFreez"])
```

```
> dadesCombi<-subset(dades2,Category=="Combi")
> dadesCooker<-subset(dades2,Category=="Cooker")
> dadesDryers<-subset(dades2,Category=="Dryers")
> dadesFreez<-subset(dades2,Category=="Freez")
> dadesRefr<-subset(dades2,Category=="Refr")
> dades3<-cbind(Combi=dadesCombi$Total,Cooker=dadesCooker$Total,Dryers=
dadesDryers$Total,Freez=dadesFreez$Total,Refr=dadesRefr$Total)
> cor<-cor(dades3);cor
>
> # Components principals
> cp<-princomp(dades3,cor=TRUE)
> summary(cp)
> cp.var<-princomp(dades3,cor=TRUE)$sdev^2
> plot(1:length(cp.var),cp.var,type="b",main="SCREE
GRAPH",xlab="Components",ylab="Variancies")
> summary(cp)$loadings
> require( FactoMineR )
> pca<-PCA(dades3) #plot.PCA(pca,choix="var",axes=1:2)
> plot.PCA(pca,choix="ind",axes=1:2)
> pca1.ind<-pca$ind$coord[,1]
> temp<-seq(2015,2017,by=1/12)
> plot(temp[-25],pca1.ind,type="l",xlab="Month",ylab="Comp.1",main="Comp.1 by
month")
> abline(lm(pca1.ind~temp[-25]),col="red")
> summary(lm(pca1.ind~temp[-25]))
> predicciocp1<-predict(lm(pca1.ind~temp[-25]),newdata=data.frame(temp=
seq(2017,2018,by=1/12)))
> plot(temp[-25],pca1.ind,type="l",xlim=c(2015,2018),ylim=c(-4,5),xlab=
"Month",ylab="Comp.1",main="Comp.1 Prediction")
> abline(lm(pca1.ind~temp[-25]),col="red",lty="dotted")
> lines(seq(2017,2018,by=1/12),prediccio,col="blue",type="l") # dibuix de les
prediccions
> abline(v=2017.04,lty="dotted",col="green") # divisio entre observacions i
prediccions
> pca2.ind<-pca$ind$coord[,2]
> temp<-seq(2015,2017,by=1/12)
> plot(temp[-25],pca2.ind,type="l",xlab="Month",ylab="Comp.2",main="Comp.2 by
month")
> abline(lm(pca2.ind~temp[-25]),col="red")
> summary(lm(pca2.ind~temp[-25]))
> predicciocp2<-predict(lm(pca2.ind~temp[-25]),newdata=data.frame(temp=
seq(2017,2018,by=1/12)))
```

```
> plot(temp[-25],pca2.ind,type="l",xlim=c(2015,2018),ylim=c(-2,2),xlab=
"Month",ylab="Comp.2",main="Comp.2 Prediction")

> abline(lm(pca2.ind~temp[-25]),col="red",lty="dotted")

> lines(seq(2017,2018,by=1/12),prediccio,col="blue",type="l") # dibuix de les
prediccions

> abline(v=2017.04,lty="dotted",col="green") # divisio entre observacions i
prediccions

>

> # ARMA

> require(tseries)

> require(forecast)

> t<-seq(2015,2017,by=1/12)

> t<-t[-25]

> # Combi

> dadesCombi2<-ts(dadesCombi$Total,start=2015,frequency=12)

> plot(dadesCombi2,type="b",xlab="Month",ylab="Total Demand",main="Combi
Demand")

> abline(lm(dadesCombi2~t),col="red")

> summary(lm(dadesCombi2~t))

> plot(dadesCombi2[1:12],type="b",xlab="Month",ylab="Total Demand",main=
"Combi Demand")

> lines(dadesCombi2[13:24],type="b",col="blue")

> legend("topright",legend=c("2015","2016"),fill=c("black","blue"),cex=0.8)

> par(mfrow=c(2,1)) # Demanda Combi

> acf(dadesCombi2,20,main="Combi Demand")

> pacf(dadesCombi2,20,main="Combi Demand")

> layout(1)

> modCombi<-auto.arima(dadesCombi2)

> summary(arma(dadesCombi2,order=c(2,0)))

> tsdiag(modCombi,gof.lag=20) ## Diagnostic (analisi de residus)

> modCombi<-arima(dadesCombi2,order=c(12,0,0))

> tsdiag(modCombi,gof.lag=20) ## Diagnostic (analisi de residus)

> ## Intervals de prediccio (al 95%)

> prediccio<-predict(modCombi,n.ahead=12) # traiem les prediccions i les
arrels del MSE

> i<-prediccio$pred-1.96*prediccio$se # extrem inferior de l'interval

> s<-prediccio$pred+1.96*prediccio$se # extrem superior de l'interval

> plot(dadesCombi2,xlim=c(2015,2018),ylim=c(0,9000),ylab="Demand",main="Combi
prediction",type="l") # dibuix de les ultimes 50 observacions

> lines(prediccio$pred,col="blue",type="l") # dibuix de les prediccions

> lines(s,col="red",lty="dashed") # dibuix de l'interval

> lines(i,col="red",lty="dashed") # dibuix de l'interval
```

```
> abline(v=2016.95,lty="dotted",col="green") # divisio entre observacions i
prediccions
> # Cooker
> dadesCooker2<-ts(dadesCooker$Total,start=2015,frequency=12)
> plot(dadesCooker2,type="b",xlab="Month",ylab="Total Demand",main="Cooker
Demand")
> abline(lm(dadesCooker2~t),col="red")
> summary(lm(dadesCooker2~t))
> dadesCooker3<-dadesCooker2-t*lm(dadesCooker2~t)$coeff[2]
> par(mfrow=c(2,1)) # Demanda Cooker
> acf(dadesCooker2,20,main="Cooker Demand")
> pacf(dadesCooker2,20,main="Cooker Demand")
> layout(1)
> par(mfrow=c(2,1)) # Demanda Cooker sense tendencia
> acf(dadesCooker3,20,main="Cooker Demand")
> pacf(dadesCooker3,20,main="Cooker Demand")
> layout(1)
> auto.arima(dadesCooker2)
> modCooker<-arima(dadesCooker2,order=c(0,0,1))
> summary(arma(dadesCooker2,order=c(0,1)))
> tsdiag(modCooker,gof.lag=20) ## Diagnostic (analisi de residus)
> ## Intervals de predicci_o (al 95%)
> prediccio<-predict(modCooker,n.ahead=12) # traiem les prediccions i les
arrels del MSE
> i<-prediccio$pred-1.96*prediccio$se # extrem inferior de l'interval
> s<-prediccio$pred+1.96*prediccio$se # extrem superior de l'interval
> plot(dadesCooker2,xlim=c(2015,2018),ylim=c(0,200),ylab="Demand",main=
"Cooker prediction",type="l") # dibuix de les ultimes 50 observacions
> lines(prediccio$pred,col="blue",type="l") # dibuix de les prediccions
> lines(s,col="red",lty="dashed") # dibuix de l'interval
> lines(i,col="red",lty="dashed") # dibuix de l'interval
> abline(v=2016.95,lty="dotted",col="green") # divisio entre observacions i
prediccions
> # Dryers
> dadesDryers2<-ts(dadesDryers$Total,start=2015,frequency=12)
> plot(dadesDryers2,type="b",xlab="Month",ylab="Total Demand",main="Dryers
Demand")
> abline(lm(dadesDryers2~t),col="red")
> summary(lm(dadesDryers2~t))
> dadesDryers3<-dadesDryers2-t*lm(dadesDryers2~t)$coeff[2]
> plot(dadesDryers2[1:12],type="b",xlab="Month",ylab="Total
Demand",main="Dryers Demand",ylim=c(0,7000))
```

```
> lines(dadesDryers2[13:24],type="b",col="blue")
> legend("bottomright",legend=c("2015","2016"),fill=c("black","blue"),
cex=0.8)
> par(mfrow=c(2,1)) # Demanda Dryers
> acf(dadesDryers2,20,main="Dryers Demand")
> pacf(dadesDryers2,20,main="Dryers Demand")
> layout(1)
> par(mfrow=c(2,1)) # Demanda Dryers sense tendencia
> acf(dadesDryers3,20,main="Dryers Demand")
> pacf(dadesDryers3,20,main="Dryers Demand")
> layout(1)
> modDryers<-auto.arima(dadesDryers2)
> summary(arma(dadesDryers2,order=c(1,1)))
> tsdiag(modDryers,gof.lag=20) ## Diagnostic (analisi de residus)
> modDryers<- arima(dadesDryers2,order=c(12,0,0))
> tsdiag(modDryers,gof.lag=20) ## Diagnostic (analisi de residus)
> ## Intervals de prediccio (al 95%)
> prediccio<-predict(modDryers,n.ahead=12) # traiem les prediccions i les
arrels del MSE
> i<-prediccio$pred-1.96*prediccio$se # extrem inferior de l'interval
> s<-prediccio$pred+1.96*prediccio$se # extrem superior de l'interval
> plot(dadesDryers2,xlim=c(2015,2018),ylim=c(0,8000),ylab="Demand",main=
"Dryers prediction",type="l") # dibuix de les ultimes 50 observacions
> lines(prediccio$pred,col="blue",type="l") # dibuix de les prediccions
> lines(s,col="red",lty="dashed") # dibuix de l'interval
> lines(i,col="red",lty="dashed") # dibuix de l'interval
> abline(v=2016.95,lty="dotted",col="green") # divisio entre observacions i
prediccions
> # Freez
> dadesFreez2<-ts(dadesFreez$Total,start=2015,frequency=12)
> plot(dadesFreez2,type="b",xlab="Month",ylab="Total Demand",main="Freezers
Demand")
> abline(lm(dadesFreez2~t),col="red")
> summary(lm(dadesFreez2~t))
> dadesFreez3<-dadesFreez2-t*lm(dadesFreez2~t)$coeff[2]
> par(mfrow=c(2,1)) # Demanda Freezers
> acf(dadesFreez2,20,main="Freez Demand")
> pacf(dadesFreez2,20,main="Freez Demand")
> layout(1)
> par(mfrow=c(2,1)) # Demanda Dryers sense tendencia
> acf(dadesFreez3,20,main="Freez Demand")
```

```
> pacf(dadesFreez3,20,main="Freez Demand")
> layout(1)
> auto.arima(dadesFreez2)
> modFreez<-arima(dadesFreez2,order=c(1,0,0))
> summary(arma(dadesFreez2,order=c(1,0)))
> tsdiag(modFreez,gof.lag=20) ## Diagn_ostic (analisi de residus)
> ## Intervals de prediccio (al 95%)
> prediccio<-predict(modFreez,n.ahead=12) # traiem les prediccions i les
arrels del MSE
> i<-prediccio$pred-1.96*prediccio$se # extrem inferior de l'interval
> s<-prediccio$pred+1.96*prediccio$se # extrem superior de l'interval
> plot(dadesFreez2,xlim=c(2015,2018),ylim=c(200,1100),ylab="Demand",main=
"Freez prediction",type="l") # dibuix de les ultimes 50 observacions
> lines(prediccio$pred,col="blue",type="l") # dibuix de les prediccions
> lines(s,col="red",lty="dashed") # dibuix de l'interval
> lines(i,col="red",lty="dashed") # dibuix de l'interval
> abline(v=2016.95,lty="dotted",col="green") # divisio entre observacions i
prediccions
> # Refr
> dadesRefr2<-ts(dadesRefr$Total,start=2015,frequency=12)
> plot(dadesRefr2,type="b",xlab="Month",ylab="Total Demand",main=
"Refrigerators Demand")
> abline(lm(dadesRefr2~t),col="red")
> summary(lm(dadesRefr2~t))
> plot(dadesRefr2[1:12],type="b",xlab="Month",ylab="Total Demand",main=
"Refrigerators Demand")
> lines(dadesRefr2[13:24],type="b",col="blue")
> legend("topright",legend=c("2015","2016"),fill=c("black","blue"),cex=0.8)
> par(mfrow=c(2,1)) # Demanda Refr
> acf(dadesRefr2,20,main="Refrigerators Demand")
> pacf(dadesRefr2,20,main="Refrigerators Demand")
> layout(1)
> modRefr<-auto.arima(dadesRefr2)
> summary(arma(dadesRefr2,order=c(1,1)))
> tsdiag(modRefr,gof.lag=20) ## Diagnostic (analisi de residus)
> modRefr<- arima(dadesRefr2,order=c(12,0,0))
> tsdiag(modRefr,gof.lag=20) ## Diagnostic (analisi de residus)
> ## Intervals de prediccio (al 95%)
> prediccio<-predict(modRefr,n.ahead=12) # traiem les prediccions i les
arrels del MSE
> i<-prediccio$pred-1.96*prediccio$se # extrem inferior de l'interval
```

```
> s<-prediccio$pred+1.96*prediccio$se # extrem superior de l'interval

> plot(dadesRefr2,xlim=c(2015,2018),ylim=c(0,7000),ylab="Demand",main=
"Refrigerators prediction",type="l") # dibuix de les ultimes 50 observacions

> lines(prediccio$pred,col="blue",type="l") # dibuix de les prediccions

> lines(s,col="red",lty="dashed") # dibuix de l'interval

> lines(i,col="red",lty="dashed") # dibuix de l'interval

> abline(v=2016.95,lty="dotted",col="green") # divisio entre observacions i
prediccions
```

Metadata

Títol del treball

Predicció de vendes amb sèries temporals

Autora

Anna Batlle Olmo

Tutor

Josep Lluís Solé

Resum

Vaig realitzar les pràctiques externes curriculars a Whirlpool S.A., empresa multinacional d'electrodomèstics. La meva funció allà era la realització de la predicció de vendes mensuals dels seus productes mitjançant la utilització dels seus mètodes propis. L'objectiu d'aquest treball és fer la predicció de vendes d'aquests mateixos productes, però amb els mètodes apressos durant els estudis de grau. En primer lloc, es realitzarà un anàlisi global de les dades, agafant els productes en conjunt i sense dividir per cap variable, per veure el seu desenvolupament al llarg del temps, i s'estudiarà el mètode de predicció utilitzat per l'empresa. A continuació, es procedirà a l'anàlisi específic per categoria de producte. En aquest apartat, s'estudiarà la possible dependència temporal de les dades. Segons els resultats obtinguts, s'intentarà trobar models que ajustin les dades i serveixin per fer les prediccions dels següents dotze mesos.

Realicé las prácticas externas curriculares en Whirlpool S.A., empresa multinacional de electrodomésticos. Mi función allí era la realización de la predicción de ventas mensuales de sus productos mediante la utilización de sus métodos propios. El objetivo de este trabajo es hacer la predicción de ventas de esos mismos productos, pero con los métodos aprendidos durante los estudios de grado. En primer lugar, se realizará un análisis global de los datos, cogiendo los productos en conjunto y sin dividir per ninguna variable, para ver su desarrollo a lo largo del tiempo, y se estudiará el método de predicción utilizado por la empresa. A continuación, se procederá al análisis específico por categoría de producto. En este apartado, se estudiará la posible dependencia temporal de los datos. Según los resultados obtenidos, se intentará encontrar modelos que ajusten los datos y sirvan para hacer las predicciones de los siguientes doce meses.

I did the external curricular practices on Whirlpool S.A, a multinational company of home appliances. My task there was to make the monthly sales prediction of their products with their own methods. The aim of this study is to make the monthly sales prediction of said products, but with the methods learned at college. First of all, a global analysis will be made, taking the whole set without splitting by any variable, in order to see its development through time, and the company's method will be studied as well. Then, we will proceed to make the specific analysis by product category. In this section, the possible temporal trend of the data will be studied. Depending on the results, we will try to find some models able to adjust the data and use them to predict the next twelve months.

Paraules clau

Sèries temporals – Predicció – Auto correlació – Tendència